NEW METHODS FOR VOICE CONVERSION

by

Oytun Türk

B.S. in Electrical and Electronics Eng., Boğaziçi University, 2000

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science
in
Electrical and Electronics Engineering

Boğaziçi University
2003

NEW METHODS FOR VOICE CONVERSION

APPROVED BY:

Assoc. Prof. Levent M. Arslan ……………………..
(Thesis Supervisor)

Assist. Prof. Engin Erzin ……………………..

Prof. Bülent Sankur ……………………..

DATE OF APPROVAL …………………………………….

ACKNOWLEDGEMENTS

I would like to thank to my thesis supervisor Assoc. Prof. Levent M. Arslan for his guidance and support. It was a great pleasure to work with him during this thesis and in all the projects we have been involved in the last three years. I would like to thank Prof. Bülent Sankur (Boğaziçi University) and Assist. Prof. Engin Erzin (Koç University) for reading my thesis and participating in my thesis committee.

I would like to thank to my family and to Aylin for their endless love, encouragement, and support. I would like to express my gratitude to Barış Bozkurt for his inspiring ideas. Special thanks go to my colleagues at Sestek Inc.

And to everyone who has participated in the subjective tests: Thanks for your patience and time. I promise, the tests will take shorter next time.

ABSTRACT

NEW METHODS FOR VOICE CONVERSION

This study focuses on various aspects of voice conversion and investigates new methods for implementing robust voice conversion systems that provide high quality output. The relevance of several spectral and temporal characteristics for perception of speaker identity is investigated using subjective tests. These characteristics include the subband based spectral content, vocal tract, pitch, duration, and energy. Two new methods based on Wavelet Transform and selective preemphasis are described for transformation of the vocal tract spectrum. A new speaker specific intonational model is developed and evaluated both in terms of accuracy and voice conversion performance. A voice conversion database in Turkish is collected and employed for the evaluation of the new methods.

ÖZET

KONUŞMACI DÖNÜŞTÜRME İÇİN YENİ YÖNTEMLER

Bu çalışma, konuşmacı dönüştürmeyle ilgili farklı yönler üzerinde yoğunlaşmakta ve yüksek kalitede çıktı sağlayan gürbüz konuşmacı dönüştürme sistemlerinde kullanılabilecek yeni yöntemlerin geliştirilmesini amaçlamaktadır. Öznel deneyler kullanılarak farklı izgel ve zamansal özelliklerin konuşmacı kimliğinin algılanmasına etkileri incelenmektedir. Bu özellikler arasında konuşma işaretlerinin farklı sıklık aralıklarındaki izgel içerikleri, gırtlak yapısı, ses perdesi, süre ve enerji yer almaktadır. Dalgacık dönüşümü ve seçici önvurgulamaya dayalı iki yeni yöntemle gırtlak yapısı dönüşümü gerçekleştirilmektedir. Konuşmacıya özgü yeni bir titremleme modeli geliştirilmiş, doğruluk ve konuşmacı dönüştürmedeki başarım açılarından incelenmiştir. Türkçe bir konuşmacı dönüştürme veri tabanı hazırlanmış ve önerilen yöntemlerin değerlendirilmesinde kullanılmıştır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AR	Auto-Regressive
ASR	Automated Speech Recognition
AVC-TTS	Adaptive Voice Conversion for Text-To-Speech Synthesis
CWT	Continuous Wavelet Transform
DFT	Discrete Fourier Transform
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transform
FD	Frequency Domain
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
Hz	Hertz
IDWT	Inverse Discrete Wavelet Transform
IPSE	Improved Power Spectrum Envelope
IVR	Interactive Voice Response
KHz	Kilo-Hertz
LF	Liljencrants-Fant
LP	Linear Prediction, Linear Predictive
LSFs	Line Spectral Frequencies
LSPs	Line Spectral Pairs
MBE	Multi-Band Excitation
MPEG	Moving Picture Experts Group
OLA	Overlap-Add
PARCOR	Partial Correlation
PCM	Pulse Code Modulation
PR	Perfect Reconstruction
PSOLA	Pitch-Synchronous Overlap-Add
QMF	Quadrature Mirror Filters

RBFN	Radial Basis Function Network
STASC	Speaker Transformation Algorithm using Segmental Codebooks
STC	Sinusoidal Transform Coding
STFT	Short Time Fourier Transform
TD	Time Domain
TTS	Text-To-Speech Synthesis
VC	Voice Conversion
VCS	Voice Conversion System
VQ	Vector Quantization

1. INTRODUCTION

1.1. Motivation

Knowledge extraction by just listening to sounds is a distinctive property and has become an important milestone in the evolution of species. Most of the animals are not only equipped with the means to extract information from the rich acoustical content of the environment and act accordingly, but they have the ability to produce sounds to interact with the environment as well. Almost all the animals with auditory perception systems are able to distinguish between the enemies, the animals that are in their troops or the animals that they can hunt by just listening. They are also capable of communicating with their environment using sound as an interaction tool. Humans have gone one step further: they have fairly advanced mechanisms that enable interaction within the species by very abstract rules of communication using voice – the language.

Perceiving the identity of others from their voices is yet another ability that only a limited number of species are known to possess. This ability is useful for heading towards the mother, avoiding the enemies, or gathering food. Human auditory system enables perception of speaker identity by just listening to a few words – in some cases even a word or a phoneme.

This study focuses on two main topics related to:

- the investigation of the abilities and properties that humans possess in perception of speaker identity
- the development and evaluation of new methods for modifying the perceived speaker identity

Recent years have witnessed the rapid advances in the speech technology with the increasing number of products which use speech as a means in human-machine interaction. This outcome was not by chance, but it was due to the efficient collaboration established between the individual speech researchers, laboratories, universities and hightech companies. The need for the improvements in human-machine interaction was even more distinctive: Humans have always been investigating new ways to ease their lives. Although the answer to the question on whether the improvements in technology always lead to improved comfort and serve for the common good is beyond the scope of this study, it is clear that technology changes the way we survive.

Naturally, speech recognition and TTS have been the priorities in research efforts directed at human-machine interaction. Effective solutions have emerged over the past years. The ways to improve naturalness in human-machine interaction is becoming an important matter of concern. Voice conversion technology will serve as a useful tool in this area because it provides new insights related to personification of speech enabled systems.

As a speech researcher, it is possible to choose from a variety of fields to conduct research: acoustical modeling, perception, recognition, synthesis, coding, linguistics etc. All these fields share many common methods and approaches. There are also several topics that serve as a connection point between these major fields, and voice conversion is absolutely one of them. It combines the methods of automated knowledge and rule extraction in speech analysis and recognition with the methods of modification and construction in speech synthesis in the light of auditory perception and linguistics.

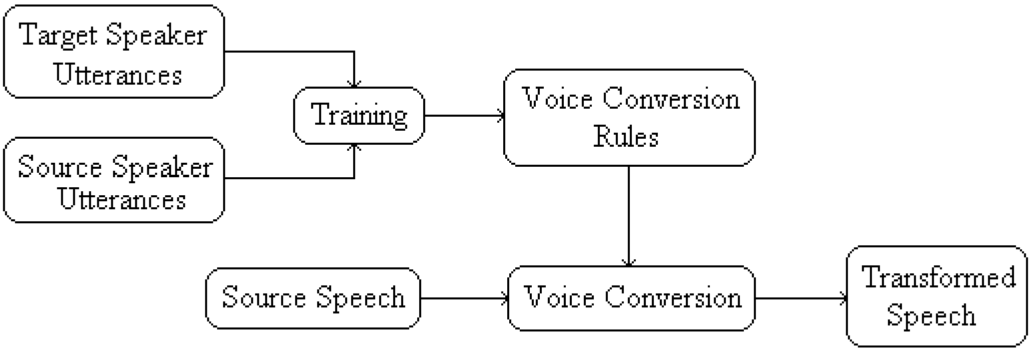


Figure 1.1. General framework for voice conversion

1.2. Definition and General Framework

Voice conversion is a method that aims to transform the input (source) speech signal such that the output (transformed) signal will be perceived as produced by another (target) speaker. A general framework for voice conversion with basic building blocks is shown in Figure 1.1.

1.3. Review of Literature

It is hard to determine an optimal method for voice conversion that can achieve success for all possible speaker characteristics and combinations. Different voice conversion systems that employ different methods exist, but at least they all share the following components:

- a method to represent the speaker specific characteristics of the speech waveform
- a method to map the source and the target acoustical spaces
- a method to modify the characteristics of the source speech using the mapping obtained in the previous step to produce converted speech

The first component is referred to as analysis or modeling, the second as training and the third as transformation. Many studies have been carried out on voice conversion in the last two decades. Examples are (Abe *et al.,* 1988), (Childers, 1995), (Baudoin and Stylianou, 1996), and (Stylianou *et al.,* 1998).

Although the details of modeling the speech waveform can be found in Chapter 3, we find it convenient to highlight major approaches here. It is common practice to model the speech waveform as a filter component driven by a source component. The filter corresponds to the vocal tract transfer characteristics which can be estimated using linear prediction (LP) methods. The parameters used in voice conversion that are extracted using LP methods include linear prediction coefficients and the parameters derived from these, such as the PARCOR coefficients (Rinscheid, 1996), the line spectral frequencies (LSFs) (Arslan and Talkin, 1997), (Arslan, 1999), (Kain and Macon, 2001), and Bark scaled LSFs (Kain and Macon, 1998a). It is also possible to approximate the vocal tract spectrum using formant frequencies, the cepstral coefficients or the sinusoidal model parameters. In this case, appropriate processing is performed to modify the relevant parameters (i.e. the formant frequencies, cepstral coefficients, or the amplitudes, frequencies and phases of the sinusoidal components). Early studies and the studies related to vocoding applications used formant frequencies for representing and modifying the vocal tract spectrum as described in (Gutierrez-Arriola et al., 1998) and (Tang et al., 2001). Sinusoidal modeling has become popular as it facilitates ease of compression and modification. In (Stylianou *et al.,* 1998), the harmonic plus noise sinusoidal model parameters are used for representing and modifying the speech signals.

The source component is usually harder to model as it contains all the remaining information in the signal such as the prosodic characteristics, lip radiation, noise, etc. Some of the problems in modeling and modifying the source component to produce natural sounding output remain unsolved. However, as the source component contains very important speaker specific information, an appropriate method to modify it for realistic voice conversion is needed. Modeling and transformation of suprasegmental characteristics such as pitch, duration and energy are are well studied and the algorithms developed provide the necessary framework for obtaining high quality output. Time Domain (TD) and Frequency Domain (FD) Pitch Synchronous Overlap-Add (PSOLA) methods (Moulines and Verhelst, 1995) are commonly used for pitch and duration scaling. The source component is also referred to as the excitation and in (Arslan, 1999) the source excitation magnitude spectrum is modified to match target speaker characteristics. It is also possible to predict the target excitation from the target training utterances as described in (Kain and Macon, 2001). Several studies investigate the relation between f0 and spectral envelope which supplies important clues in modifying the excitation spectrum: (Tanaka and Abe, 1997), and (Kain and Stylianou, 2000).

The problem of estimating the correspondence between the source and the target acoustical spaces is in fact a learning problem. It is well studied in the machine learning and artificial intelligence literature. Principal learning methods were successfully applied for the purpose of training in voice conversion such as Vector Quantization (VQ), Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), Artificial Neural

Networks (ANNs) and Radial Basis Function Networks (RBFNs). The main steps in training usually involve :

- Pre-processing and proper alignment of source and target training data
- Choice of a speech model that will capture important characteristics of the speakers
- Analysis of the training data for estimating the speech model parameters
- Employing a learning method which will automatically estimate a mapping between the source and target models and is able to generalize to unobserved data successfully

In order to obtain the corresponding acoustical events between the source and the target, acoustical alignment should be carried out to determine the event boundaries within the utterances. The most straightforward way is to manually mark the corresponding acoustical events on the source and target recordings and to select the acoustical parameters corresponding to the current acoustical event. However, it is much time saving to use automatic methods for alignment before obtaining the mapping function. Dynamic Time Warping (DTW) was the former approach for alignment (Itakura, 1975). After the development of Hidden Markov Modeling techniques in speech processing, HMMs have been widely used for automatic alignment instead of DTW. Examples are (Kim et al., 1997), (Pellom and Hansen, 1997), and (Arslan, 1999). Automatic phoneme recognition can be an alternative (Kain and Macon, 1998a), however this will make the system language dependent.

The mapping function can be obtained using vector quantization as in (Rinscheid, 1996) using a self organizing map, or as in (Hashimoto and Higuchi, 1996) using vector field smoothing. The spectral vectors are usually kept in codebooks which correspond to a one-to-one mapping between the discrete source and target acoustical space. Previous research on voice conversion employing codebook based methods include (Abe *et al.,* 1988), (Baudoin and Stylianou, 1996), and (Arslan, 1999). In this case, interpolation is necessary for converting source vectors that are not in the codebook to account for the unobserved data. An alternative for avoiding interpolation is to use continuous transformation functions such as Gaussian Mixture Models (Stylianou *et al.*, 1995), (Stylianou and Cappe, 1998), (Kain and Stylianou, 2000), and (Kain, 2001). In

(Gutierrez-Arriola et al., 1998), source to target mapping is learned through a set of linear regression rules. Artificial Neural Networks and Radial Basis Function Networks (Watanabe *et al.*, 2002) are yet other alternatives for estimating the mapping between the source and the target parameter sets.

Once training has been completed, the voice conversion system has gathered sufficient information to transform any source speech signal into the target's voice in the transformation stage. This stage employs different methods to modify source speaker characteristics in order to obtain an output that sounds as close to the target speaker's voice as possible. The nature of these methods heavily depend on the speech model being used. Appropriate modification of source parameters and re-synthesis using the modified parameters produces the output (transformed) speech. Most voice conversion systems employ the following steps in transformation:

- Estimation of the speech model parameters from input speech
- Modification of input speech by modifying model parameters using the knowledge extracted during training
- Synthesis of the output speech from the modified parameters

1.4. Applications

Voice conversion will serve as an invaluable tool for many applications in speech technology. The following sub-sections include applications of voice conversion with relevant references and new applications as demonstrated by this study.

1.4.1. Dubbing Applications

Voice conversion can be used for looping and dubbing applications as we describe in (Turk and Arslan, 2002). Looping is defined as replacing the undesired utterances in a speech recording by desired ones. This method can be used for processing movies for TV broadcast. In order to obtain transparent quality such that the listeners will not be able to distinguish the replacement necessiates the use of voice conversion. Following dubbing applications are possible with voice conversion technology:

- Dubbing the whole movie by using only several dubbers
- Regenerating the voices of actresses/actors who are not alive or who have lost their voice characteristics due to old age
- Generating the voice of famous actresses/actors in another language which they do not know
- Dubbing in radio broadcasts

All these applications require high quality output. It is also important that the methods to be used must facilitate fast and convincing voice conversion. As an example for a dubbing application, we may consider a movie which is originally in English and which will be translated and dubbed in Turkish. When conventional dubbing methods are used, the English text is translated to Turkish. Dubbers perform dubbing using the translated text. Thus, the speaker identity of the actor in the original soundtrack is lost. However, with voice conversion technology, a famous actor can talk in a language that he does not know. The following steps provide the outline of the application:

- The text in English should be read and recorded by a dubber who speaks both Turkish and English fluently.
- The voice conversion system should be trained using the recordings of the dubber as the source and the original recordings in English as the target.
- The dubber should record the corresponding Turkish text.
- Turkish recordings of the dubber should be transformed to the voice of the actor / actresses employing the information obtained in the training phase.

Voice conversion can also be used in the applications in which the voice of a particular speaker should be produced on the fly. As an example, consider TV/radio broadcast flows. Momentary changes in the flow necessiates the speaker to record the changes immediately, so s/he must be available in the course of the broadcast. But if voice conversion is employed, the speaker should appear all at once, to record the training utterances. The voice of another speaker can be converted to the original speaker’s voice when s/he is not available. Even better solutions will emerge as voice conversion systems improve: you can pay for the voice of a particular speaker once, and then use that voice font forever, even after the speaker passes by.

1.4.2. Speech Processing Applications

Text-To-Speech Synthesis (TTS) has become an important tool in the systems that facilitate human-machine interaction. This is due to the fact that the quality of synthesis output has reached a level that is suitable for many practical applications. Although people have not been very comfortable with synthetic voice in the past years, the advances in TTS database design, development of new methods for selection from a large speech corpora, and advances in signal processing and language modeling produced several TTS systems that provide fairly high quality synthesis. E-mail readers, Interactive Voice Response (IVR) systems and screen readers for the blind have become the typical application areas for synthesized speech.

Typically, TTS systems can generate speech by only a few speakers. The reason for synthetic voice being restricted to a few speakers is related to the cost of employing more speakers for synthesis because a separate database should be recorded, segmented and processed for each speaker. In fact, the performance of any synthesis system is also strictly dependent on the database and on the voice that is being used. Some voices are easier to analyze and produce better results when used in synthesis (Syrdal *et al.*, 1998). It takes several days to several weeks to design the TTS database. On the other hand, voice conversion is very economical in the sense that the database to be handled is significantly shorter (several minutes) than an ordinary TTS database. Voice conversion can be used to generate new TTS voices without the need to generate and process a separate TTS database for each speaker. We can even have the following scenario possible: an e-mail message can be read by the sender's voice provided that he/she has recorded a voice conversion database once before. Several studies have addressed the problem of using voice conversion in TTS such as (Ribeiro and Trancoso, 1997), (Kain and Macon, 1998b), and (Kain and Macon, 1998c).

Multilingual TTS provides the framework necessary for fast and reliable communication across the national borders. Several multilingual TTS sytems are in the market. While the number of languages that a TTS system can speak and the quality of the output are the major concerns of TTS research, an important question remains unattended: How can we have an adaptive TTS system that can produce speech in another speaker's voice? Of course, the straight answer is to create a separate database for each speaker. This method is employed in all TTS systems. However, this is unrealistic when the aim is to be able to generate synthetic voice with any user's voice for real-time applications. We need a tool to modify the TTS database or the TTS output for modifying the perceived speaker identity, and the natural tool is voice conversion. In fact, we refer to the method of modifying TTS databases and/or outputs using voice conversion as Adaptive Voice Conversion for Text-To-Speech Synthesis (AVC-TTS). Synthesis of speech with any user's voice will be possible with this method.

Adaptive voice conversion systems might also be used in applications related to the healthcare industry. Patients of throat cancer and people with severe voice disorders could benefit from using voice-adaptive TTS systems. These people can record a voice conversion database in the early stages of their illness. This database can be used in the future to reproduce the voice of the patient in the case of a voice disorder or loss.

E-mail readers serve as an important tool in Interactive Voice Response (IVR) systems. With these systems, people can listen to their e-mail messages on the phone. Personification of e-mail readers using voice conversion will provide the possibility to attach a voice font to each personality and the messages can be read by the sender's voice or any voice the user may prefer. Voice disguise or modification is another application.

Speech enhancement methods are widely used for the reconstruction of old recordings. Noise and unintelligible parts exist in these recordings due to problems in the recording technology and the recording environment. As an example, magnetic tapes are subject to corruption by time. However, it is impossible to restore a recording which is completely corrupted or the noise level is beyond a limit. If we have sufficient amount of clean recordings of the person whose voice we are trying to produce, we can generate good quality recordings using voice conversion.

Automated Speech Recognition (ASR) technology aims to implement computer systems that are able to recognize speech. Training these systems with voices of many speakers is necessary for robustness. However, collecting and processing SR databases is a tedious task. It is possible to use voice conversion in the recognition phase for reducing speaker variability instead of training the system with many voices.

Speaker verification and identification systems are used in security and forensic applications for automated retrieval of identity from speech signals. The know-how obtained in voice conversion will serve as an important information source for such systems as the most important characteristics for perception of speaker identity will be determined.

1.4.3. Multimedia/Music Applications

Voice conversion techniques can be successfully applied in multimedia and musical applications. An example is the Karaoke machine. An ordinary voice can be transformed into a famous singer's voice. A study on the use of voice modification for Karaoke is described in (Verhelst *et al.*, 2002). However, in that study only appropriate pitch scale and time scale modifications are carried out in a time warping framework and the identity of the Karaoke singer is left unchanged. Generation of virtual voices for virtual characters created with 3D face synthesis for animations and movies is another application.

As musical applications are considered, pitch modification while preserving the spectral identity of the musical instrument is possible as described in (Drioli, 1999). In this case, the problem is to estimate the spectral envelope while modifying the pitch, because the spectral envelope changes with pitch in musical instruments. Voice conversion methods are applied for generating the spectral envelope at the desired pitch value. One can also train such a system with two different instruments and convert the spectral envelope as well while modifying the pitch. This method will be very useful for synthesizers.

2. PROBLEM STATEMENT

Voice conversion is a fertile field for speech research as the problems of concern are related to almost all of the primary topics in speech processing. First, the analysis stage of voice conversion is related to developing appropriate models that capture speaker specific information and estimating the model parameters which are closely related to acoustical modeling, speech coding, and psychoacoustics. Next, the relation between the source and target models must be determined and generalized to unobserved data. The learning and generalization processes relate voice conversion with speech/pattern recognition, and machine learning. Finally, convenient methods must be employed for processing the source signal with minimized distortion and maximized resemblence of the output to the target speaker. These methods are also addressed in speech synthesis and coding applications. Robustness is perhaps the most important point of concern in voice conversion as the aim is to develop methods that perform well for a wide variety of source-target speaker pairs. In this study, we explore new methods related to all three dimensions described above- analysis, learning and synthesis, with the aim of developing a robust and automated system that minimizes the need for user interference. This research is far from being a complete solution to all the problems related to voice conversion, however it provides new insights and solutions.

Subjective testing is the most realistic method for both the investigation of the characteristics that human auditory system possesses and the performance evaluation of the voice conversion systems. During this study we have designed several subjective evaluation methods for both assessing the importance of different characteristics for perception of speaker identity and evaluation of the new methods we have proposed. In the preliminary tests, we first evaluate the importance of different frequency bands for perception of speaker identity. We study the relevant features for voice conversion: the vocal tract, pitch contour, phonemic durations and energy contour. A subjective test is designed for evaluating all possible combinations of these features in perception of speaker identity.

Next, we investigate the problem of vocal tract modeling, which is one of the most important speaker specific characteristic in voice conversion. We describe new methods for detailed estimation and modification of the vocal tract spectrum. We address the problems at high sampling rates for obtaining high quality output. In general, we adapt a subband based framework taking into account perceptual characteristics of the human auditory system. Two new methods for modeling and transforming the vocal tract are proposed. The first method describes a subband based system that relies on Discrete Wavelet Transform (DWT) for subband decomposition and reconstruction. The second method, selective preemphasis, provides the means for detailed spectral envelope estimation and for modification of spectral resolution in different subbands. Modification of the speech prosody will be another point of concern. The evolution of the fundamental frequency values over time produces the intonational characteristics which is an important clue for the perception of speaker identity as the subjective tests demonstrate. For this reason, we investigate a new method for detailed pitch contour transformation.

We describe the design of a voice conversion database to be used in the tests. A subjective test is designed for evaluating the new methods developed for voice conversion. DWT, selective preemphasis and pitch transformation methods are compared in this subjective test. We have carried out objective analysis on the performance of two spectral estimation methods. We have also investigated the objective performance of the new voice conversion methods.

Finally, a software tool, VOX, is implementd for fast and reliable voice conversion. This software tool is also referred to as a Voice Conversion System (VCS). It integrates all necessary tools for voice conversion in a single interface: tools for waveform analysis, recording, training, transformation, subjective and objective testing. It enables people who are not speech processing experts to perform voice conversion in a fast, reliable manner offering high quality output with an ordinary personal computer (Please refer to Appendix A for more detail).

3. THEORETICAL BACKGROUND

3.1. Theory of Speech Production

Speech production process is initiated by the air flow generated by the lungs. The air flow passes through the larynx which contains vocal folds (cords). The space between the vocal folds is known as the glottis. In the voiced sounds like /a/ and /e/, the air flow causes the vocal folds to vibrate and produce a quasi-periodic glottal waveform. For the unvoiced sounds like /s/ and /f/, the vocal folds are open and the source component contains noise-like spectral energy distribution.

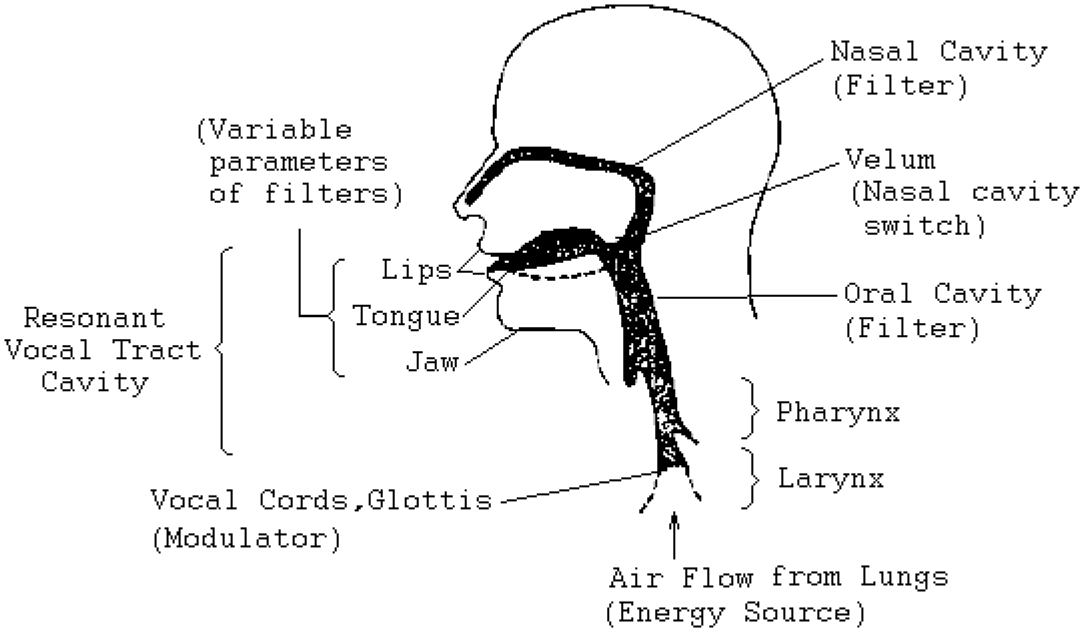


Figure 3.1. Human speech production system (Wu, 2003)

The spectrum of the source component is shaped further by the cavities in the vocal tract that reside in the pharynx, in the oral and nasal areas. Also the tongue and the lips modify the output. The characteristics of the filtering action does not change very fast in general, so it is possible to estimate the filter parameters from a short speech segment of typically 10 ms - 40 ms. length. Human speech production system is shown in Figure 3.1.

When examined on a short-time basis, the speech waveform may exhibit different characteristics. As an example, if the vocal folds vibrate as explained above, voiced sounds are produced. The glottal waveforms for unvoiced sounds have rather smooth spectra that can be successfully approximated by White Gaussian Noise. The speech signal has a sudden (impulse-like) change in the energy for plosive sounds like /p/ and /t/. The effect of the nasal cavities is much more dominant for nasalized sounds like /n/ and /m/ for which the spectra possess spectral nulls as well as spectral peaks.

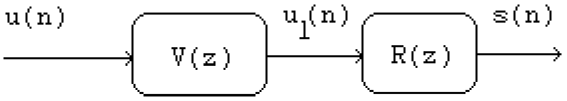


Figure 3.2. Speech production model

The observations on different type of sounds that the human speech production mechanism is able to produce have led to a generalized model of speech production: The speech waveform is modeled as the output of a time-varying all-pole filter driven by the source component. The source component is the glottal waveform, noise or a mixture of two. This model is known as the source/filter model of speech production. Its flowchart is given in Figure 3.2. Note that u(n) denotes the source signal and it is shaped by the vocal tract transfer function V(z) to produce $u_l(n)$, the volume velocity at the lips. The lip radiation filter, R(z), shapes the signal spectrum further. R(z) can be approximated by the delay term $1\text{-}z^{-1}$. Finally, s(n) is the sound pressure at the microphone. The transfer function of the vocal tract filter can be approximated as the transfer function of an all-pole filter (Equation 3.1).

$$V(z) = \frac{Gz^{-\frac{P}{2}}}{1 - \sum_{j=1}^{P} a_j z^{-j}} = \frac{Gz^{-\frac{P}{2}}}{A(z)} \cong \frac{G}{A(z)} \tag{3.1}$$

The delay term $z^{-P/2}$ in the numerator of V(z) is usually neglected. The filter coefficients, a_j's, can be estimated using linear prediction methods. In fact, there are several methods for estimating the parameters of the general scheme described above. In Sections 3.2 and 3.3, we will cover several common methods for modeling the filter and the source component. The linear predictive methods will be in the focus of the descriptions as we have extensively used it for acoustical modeling in the voice conversion methods proposed.

3.2. Modeling the Filter Component

3.2.1. Linear Prediction Coefficients

As the vocal tract transfer function V(z) does not change much during the impulse response of R(z), we can use the model shown in Figure 3.3 instead of the model described above.

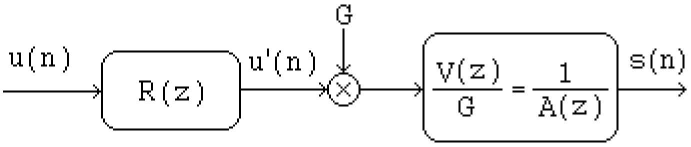


Figure 3.3. Modified source/filter model for speech production

$$s(n) = Gu^{'}(n) + \sum_{j=1}^{P} a_j s(n-j) \tag{3.2}$$

The speech waveform s(n) is calculated by Equation 3.2. If the gains of the vocal tract resonances are high, the second term dominates in the calculation, so the speech waveform can be approximated as:

$$s(n) \cong \sum_{j=1}^{P} a_j s(n-j) \tag{3.3}$$

Equation 3.3 clarifies the reason for referring to this analysis method as linear prediction: we predict (or estimate) the current sample of the speech waveform as a linear combination of the past samples. The prediction error is given by Equation 3.4 in the time domain and by Equation 3.5 in the z-domain. The aim of linear prediction (LP) analysis is to find a set of LP coefficients, a_j's, that minimize the sum of squared errors for an input speech segment. The values of a_j's that minimize $Q_E = \Sigma e^2(n)$, where n is the index of the samples at the current frame is given in matrix-vector form by Equation 3.6.

$$e(n) = s(n) - \sum_{j=1}^{P} a_j s(n-j) = s(n) - a_1 s(n-1) - a_2 s(n-2) - \\ \ldots - a_P s(n-P) \tag{3.4}$$

$$E(z) = S(z)A(z) \tag{3.5}$$

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = \mathbf{\Phi}^{-1}\mathbf{c} \qquad \text{where} \qquad \phi_{i,j} = \sum s(n-i)s(n-j) \qquad \text{and} \qquad \mathbf{c}{=}\phi_{\mathbf{i},\mathbf{0}} \tag{3.6}$$

Note that, **Φ** is symmetric and positive semi-definite. If we can find a way to make **Φ** Toeplitz, we have considerable reduction in computation time and improved accuracy in finding the inverse of matrix **Φ**. Note that the number of operations for the inversion of a PxP matrix is on the order of P^3. However, if the matrix under consideration is Toeplitz (i.e. has constant diagonals), inversion requires on the order of P^2 operations which yields a considerable amount of reduction in computation. If we use windowing and calculate $\phi_{i,j} = \Sigma s(n-i)s(n-j)$ as an infinite sum in the range (-∞,+∞), we have:

$$\phi_{i,j} = \phi_{|i-j|,0} = R_{|i-j|} = R_k \tag{3.7}$$

where R_k is the autocorrelation sequence of the windowed speech waveform. The matrix **Φ** becomes Toeplitz in this case. The set of equations to be solved are known as the Yule-Walker Equations (Equation 3.8). Inversion procedure to find the LP coefficients **a** is known as the Levinson-Durbin algorithm. It requires on the order of P^2 operations.

$$\mathbf{\Phi a} = \mathbf{c} \tag{3.8}$$

It is also possible to obtain a_j's using the covariance method by calculating $\phi_{i,j} = \Sigma s(n-i)s(n-j)$ for n=0,..., N-1 without windowing. In this case, the number of required operations increase as **Φ** is no longer Toeplitz. The covariance method may result in an unstable filter V(z), so the filter must be checked for stability. The poles outside the unit circle should be reflected to force stability. Checking the pole stability and reflecting the poles that cause unstability require extra computational effort.

Moreover, the covariance method is more sensitive to the precise position of the speech frame in relation to the vocal fold closure instants. The advantages of the covariance method are:

- Windowing is not required
- Infinite spectral resolution is obtained
- Gives better results than the autocorrelation method

The method of estimating the filter coefficients depend on the requirements of the application at hand. After the LP coefficients are estimated, the vocal tract filter V(z) is given by P+1 parameters :

$$V(z) = \frac{G}{1 - \sum_{j=1}^{P} a_j z^{-j}} \tag{3.9}$$

where a_j's are the LP coefficients, P is the prediction order, and G is the gain. In the autocorrelation approach G can be estimated as in Equation 3.10. Note that the LP coefficients are also referred to as AR coefficients because if we consider the input to the vocal tract filter as a random variable, the speech production model given above corresponds to an auto-regressive (AR) process of order P.

$$G^2 = R_n(0) - \sum_{j=1}^{P} a_j R_n(j) \tag{3.10}$$

The matrix **Φ** is always non-singular but a measure of singularity provides insights to improve the numerical properties of the LP analysis procedure. This measure is known as the condition number and it is given by the ratio of the largest eigenvalue to the smallest eigenvalue of the matrix under consideration. For large prediction orders, the condition number of **Φ** tends to the ratio $S_{max}(w)/S_{min}(w)$. So the numerical properties of the LP analysis procedure can be improved by applying a spectral flattening filter prior to analysis. This method is known as preemphasis and the spectral fall-off can be compensated with a 1st order high-pass filter with a zero near z = 1:

$$P(z) = 1 - \alpha z^{-1} \tag{3.11}$$

From a spectral flatness point of view, the optimum value for α is ϕ_{10}/ϕ_{00}. This value can be obtained by an LP analysis of order 1 prior to the actual analysis. The method of adaptively modifying α for each speech frame is known as adaptive preemphasis. Note that, P(z) is approximately a differentiator with a normalized corner frequency of approximately $(1-\alpha)/2\pi$. The corner frequency is typically placed in the range 0-150 Hz in the applications.

3.2.2. Alternative Parameter Sets

Several alternative parameter sets derived from the LP coefficients have been proposed for different applications. Particularly, the LP coefficients do not possess desired characteristics for interpolation and quantization. Coding applications require good quantization characteristics and synthesis applications require good interpolation characteristics. The disadvantages of LP coefficients are:

- Stability check problem: It is not easy to verify that a given LP coefficient set represents a stable filter.
- Quantization problem: The frequency response of the vocal tract filter is sensitive to changes in the LP coefficients.
- Interpolation problem: Interpolating two stable LP coefficient sets does not produce a smoothly modified version of the vocal tract frequency response. Stability is not even guaranteed.

There are two parameter sets that are commonly used in applications: the cepstral coefficients and the line spectral frequencies (LSFs). These alternative representations exhibit different characteristics as explained in the following sections.

3.2.3. Cepstral Coefficients

The cepstral coefficients are extensively used in speech recognition because they posses well discrimination characteristics between phonemes. It is possible to approximate the cepstral coefficients using Gaussian distributions within a phoneme. The term cepstrum is defined as the Inverse Fourier Transform of the log spectrum:

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \log V(e^{jw})e^{jwn}\,dw \tag{3.12}$$

The cepstral coefficients, c_n's, can be obtained directly from the LP coefficients as follows:

$$c_n = a_n + \frac{1}{n}\sum_{j=1}^{\min P,n-1}(n-j)c_{n-j}a_j \tag{3.13}$$

The resulting coefficients are called the *complex cepstrum coefficients* although they are real. The *cepstrum coefficients* are calculated by a slight modification in the derivation of c_n by using $\log|V(e^{jw})|$ instead of $\log(V(e^{jw}))$. In this case, the coefficients equal to the half of the c_n's given by Equation 3.13 except that c_0 remains the same. Note that stability check is also difficult for the cepstral coefficients.

3.2.4. Line Spectral Frequencies (LSFs)

The LSFs are calculated using a symmetric and an anti-symmetric polynomial obtained from A(z). The symmetric polynomial, P(z), and the anti-symmetric polynomial Q(z) are as follows:

$$P(z) = A(z) + z^{-(P+1)}A^*(z^{*-1}) = 1-(a_1+a_P)z^{-1} - (a_2+a_{P-1})z^{-2} - \dots - (a_P+a_1)z^{-P} + z^{-(P+1)} \tag{3.14}$$

$$Q(z) = A(z) - z^{-(P+1)}A^*(z^{*-1}) = 1-(a_1-a_P)z^{-1} - (a_2-a_{P-1})z^{-2} - \dots - (a_P-a_1)z^{-P} - z^{-(P+1)} \tag{3.15}$$

The vocal tract tansfer function V(z) is stable if and only if all the roots of P(z) and Q(z) are on the unit circle and they are interleaved. We proove that P(z) and Q(z) satisfy these conditions below. If the roots of P(z) are at $\exp(2\pi j f_i)$ for $i = 1,3,...$ and those of Q(z) are at $\exp(2\pi j f_i)$ for $i = 0, 2, ...$ with $f_{i+1} > f_i \geq 0$ then the LSF frequencies are defined as $f_1, f_2, ..., f_p$. Note that $f_0 = +1$ and $f_{p+1} = -1$.

Theorem: All the roots of P(z) and Q(z) lie on the unit circle

Proof: The roots of P(z) and Q(z) can be estimated by setting these polynomials equal to zero as follows:

$$P(z) = 0 \Rightarrow A(z) = -z^{-(p+1)}A^*(z^{*-1}) \Rightarrow H(z) = -1 \tag{3.16}$$

$$Q(z) = 0 \Rightarrow A(z) = +z^{-(p+1)}A^*(z^{*-1}) \Rightarrow H(z) = +1 \tag{3.17}$$

where

$$H(z) = \frac{A(z)}{z^{-(p+1)}A^*(z^{*-1})} = z\prod_{i=1}^{P}\frac{1-x_iz^{-1}}{z^{-1}(1-x_i^*z)} = z\prod_{i=1}^{P}\frac{z-x_i}{(1-x_i^*z)} \tag{3.18}$$

x_i's are the roots of A(z). If all x_i's lie inside the unit circle, the absolute values of the terms making up H(z) are either all greater than 1 or all less than 1. Calculating the absolute value of a typical term, we get:

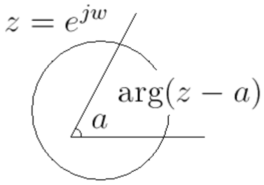
$$\begin{aligned}
\mid \frac{z-x_i}{1-x_i^*z} \mid > 1 &\Rightarrow \mid 1-x_i^*z \mid < \mid z-x_i \mid \\
&\Rightarrow (1-x_i^*z)(1-x_i^*z)^* < (z-x_i)(z-x_i)^* \\
&\Rightarrow (1-x_i^*z)(1-x_iz^*) < (z-x_i)(z^*-x_i^*) \\
&\Rightarrow 1-x_i^*z-x_iz^*+x_ix_i^*zz^* < zz^*-x_i^*z-x_iz^*+x_ix_i^* \\
&\Rightarrow 1-x_ix_i^*-zz^*+x_ix_i^*zz^* < 0 \\
&\Rightarrow (1-\mid x_i \mid^2)(1-\mid z \mid^2) < 0 \\
&\Rightarrow \mid z \mid > 1 \;\; since \mid x_i \mid < 1
\end{aligned} \tag{3.19}$$

So, each term is greater than or less than 1 according to |z|. If |z|>1 then each term is greater than 1 and vice versa. So |H(z)|=1 if and only if |z|=1. So the roots of P(z) and Q(z) lie on the unit circle.

Theorem: The roots of P(z) and Q(z) are interleaved.

Proof: We want to find the values of $z = e^{jw}$ that make H(z) = +1 or H(z) = -1. This is equivalent to finding the values that make arg(H(z)) a multiple of π. If $z = e^{jw}$, then:

$$\begin{aligned}\arg(H(e^{jw})) &= \arg(e^{j(1-P)w}\prod_{i=1}^{P}\frac{(e^{jw}-x_i)}{(e^{-jw}-x_i^*)}) \\ &= (1-P)w+\sum_{i=1}^{P}(\arg(e^{jw}-x_i)-\arg(e^{-jw}-x_i^*)) \\ &= (1-P)w+2\sum_{i=1}^{P}(\arg(e^{jw}-x_i)) \qquad (3.20)\end{aligned}$$



As w goes from 0 to 2π, arg(z-a) changes by +2π if |a|<1. Therefore, as w goes from 0 to 2π, arg(H(e^{jw})) increases by (1-P)2π + 2P.2π = (1+P)2π. Since H(e^{jw}) goes round the unit circle (1+P) times, it must pass through each of the points +1 and –1 at least (1+P) times. Since P(z) and Q(z) are (P+1)th order polynomials, they have only P+1 roots. So H(e^{jw}) can not pass through either +1 or –1 more than P+1 times. It follows that H(e^{jw}) passes through +1 and –1 alternately exactly P+1 times each. Note that arg(H(z)) varies most rapidly when z is near one of the x_i, so the LSFs will cluster near the formants.

The LSFs have the following useful properties:

- Stability check is easy. If the LSFs are in ascending order in the range [0,1], the resulting filter is guaranteed to be stable.
- Interpolation is possible.
- As the LSFs are strongly correlated with each other, they can be quantized efficiently.
- When two LSF values are close to each other, a spectral peak is likely to occur between them which is useful for tracking formants and spectral peaks.
- It is easy to adapt a perceptual representation such as the Bark scale because the LSFs are pure frequency values.

The disadvantage of LSFs is the requirement to calculate the roots of P(z) and Q(z) polynomials which correspond to solving polynomials of order P. When the sampling rate is high, the sufficient prediction order will be high and the performance of the root-finding algorithms will degrade. As an example, in (Rothweiler, 1999), the proposed root-finding method works well for orders up to 24. In fact, this issue inspired our research on investigating new spectral estimation techniques at high sampling rates using lower prediction orders.

3.2.5. Sinusoidal Model

In fact the sinusoidal model is a complete model of the speech production mechanism (McAulay and Quatieri, 1995). The excitation signal is represented as a linear combination of sinusoids and passing this excitation signal through the vocal tract filter results in the sinusoidal representation of the speech wavefrom as given by Equation 3.21. The amplitudes of the sinusoids, A_l's, are estimated using the peaks of the DFT spectrum and considering the harmonics of the neighbouring frames.

$$s(n) = \sum_{l=1}^{L} A_l \cos(w_l n + \phi_l) \tag{3.21}$$

In the analysis stage, the parameters of the model are estimated frame-by-frame. Either these parameters or coded/modified versions of them are used in synthesis to output the kth synthetic frame as follows:

$$\hat{s}^k(n) = \sum_{l=1}^{L^k} A_l^k \cos(w_l^k n + \phi_l^k) \tag{3.22}$$

3.2.6. Improved Power Spectrum Envelope (IPSE) Analysis

IPSE analysis is a spectrum envelope extraction method proposed in (Tanaka and Abe, 1997). The aim is to extract the spectral envelope pitch-synchronously for estimating the vocal tract spectrum in detail. The algorithm makes use of the spectral peaks and f0 value in the estimation. The main steps are as follows:

- A speech frame of 2 to 5 pitch periods is windowed using a Hamming window.
- The log-power spectrum is calculated by FFT.
- The local-maximum value of the log-power spectrum is sampled at f_n intervals ($nf_o - f_o / 2 < f_n < nf_o + f_o / 2$, where n is an integer).
- If the interval between f_n and f_{n+1} is larger than 1.5 times f_o, the local peaks of the log-power spectrum within the interval are added to the sequence obtained above.
- The samples are linearly interpolated and resampled at f_o/n intervals where n is the integer that gives the maximum value of f_o/n while $f_o/n < 50$ Hz.

- The resampled lines are approximated by the following cosine model:

$$Y(\lambda) = \sum_{i=0}^{M} A_i \cos(i\lambda) \quad for \quad 0 \leq \lambda \leq \pi \tag{3.23}$$

However, we have used cubic spline interpolation for simplicity in the following examples. This type of interpolation also matches the spectrum well. In the following figures, we compare the performance of the LP analysis with the IPSE method for the estimation of the spectral envelope. It is clear that IPSE tracks the spectral peaks better than the LP method.

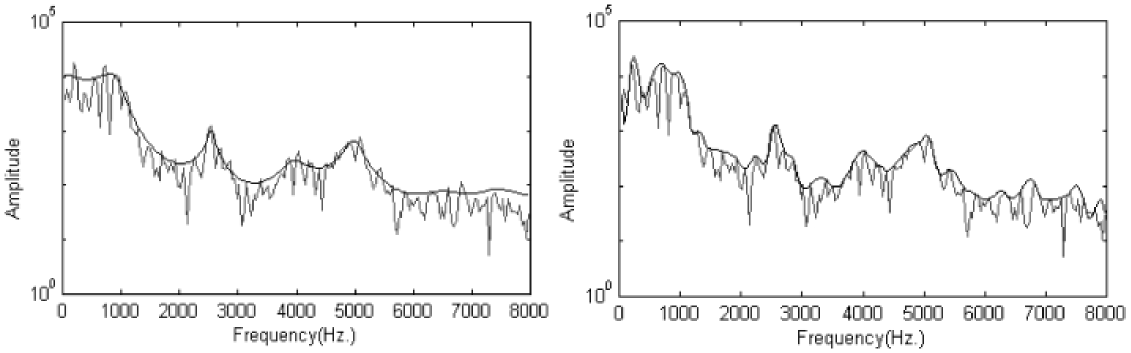


Figure 3.4. Spectral envelope obtained by LP analysis (left) and IPSE method (right). (Prediction order was 18 for a sampling rate of 16 KHz.)

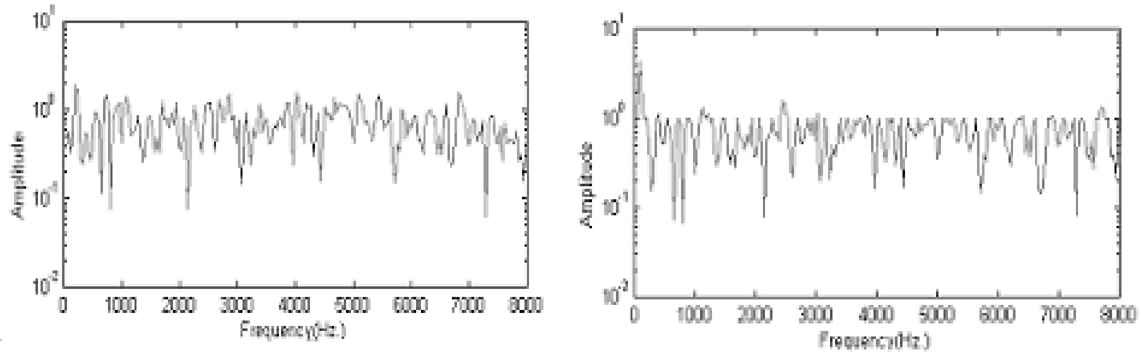


Figure 3.5. Excitation spectrum after the LP analysis (left)
and the IPSE method (right)

As the IPSE method tracks the spectral envelope in more detail than LP analysis (Figure 3.4), the resulting excitation spectrum is smoother as shown in Figure 3.5.

3.3. Modeling the Source Component

We can estimate the source component by applying inverse filtering techniques on the speech signals. The method relies on estimating the vocal tract filter and extracting the source component using the estimated filter coefficients frame-by-frame. Figure 3.6 shows a simple inverse-filtering algorithm output. Most algorithms output the derivative of the glottal flow for voiced sounds. The noise-like components in the voiced source signal are usually filtered out by a lowpass filter.

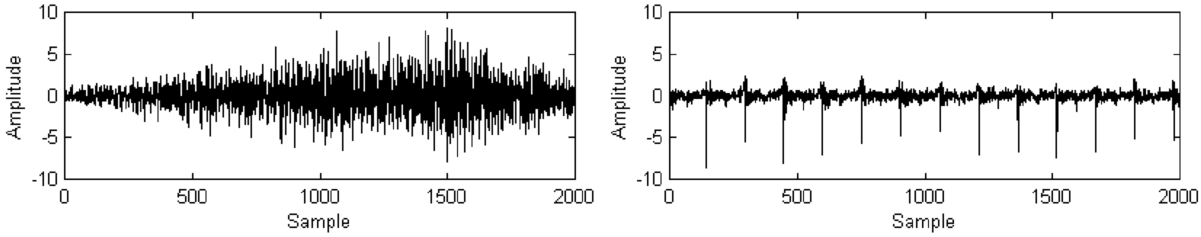


Figure 3.6. Glottal waveforms for unvoiced (left) and voiced (right) signal segments

3.3.1. Impulse/Noise Model

The simplest model for the source component relies on voiced/unvoiced decisions. For voiced sounds like /a/, /e/ we observe a periodic pattern in the speech waveform as explained above. In this case, the source component can be approximated as impulses located according to the pitch period. For the unvoiced sounds like /s/, /sh/ the spectrum of the source component is well approximated by white noise spectrum.

The flowchart for a simple LPC vocoder is shown in Figure 3.7. In LPC vocoders, the vocal tract filter is excited with either impulses or noise in the synthesis stage. The impulse assumption in voiced segments is an approximation and the output of LP based synthesizers is buzzy.

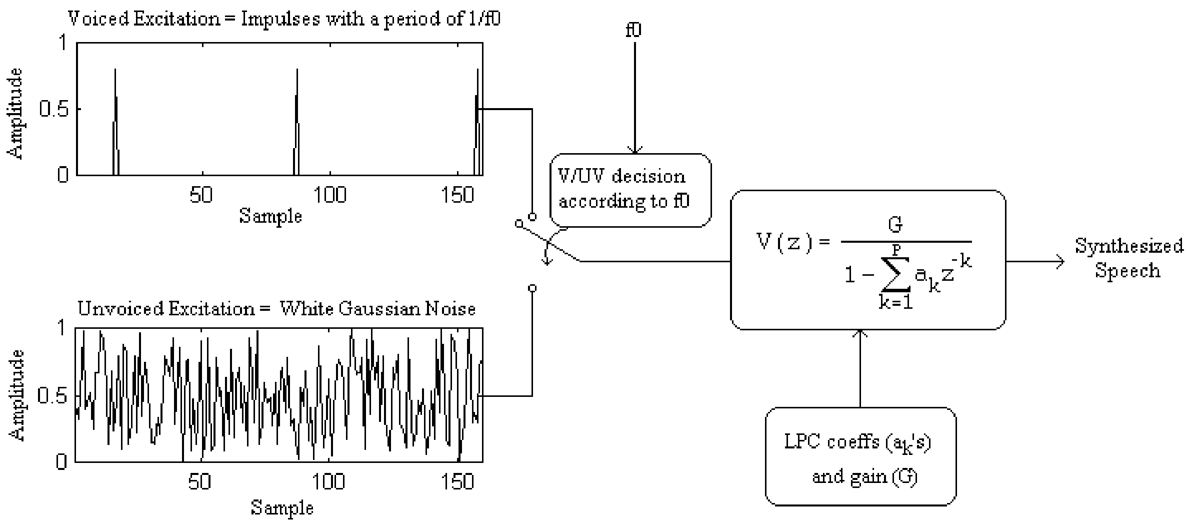

Voiced Excitation = Impulses with a period of 1/f0
Amplitude
1
0.5
0
50
100
150
Sample
Unvoiced Excitation = White Gaussian Noise
Amplitude
1
0.5
0
50
100
150
Sample
f0
V/UV decision according to f0
$V(z) = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}}$
Synthesized Speech
LPC coeffs (a_k's) and gain (G)

Figure 3.7. LPC vocoder flowchart

3.3.2. Multiband Excitation (MBE) Model

This model is a generalization of the impulse/noise model. The excitation spectrum is processed in different subbands and a voicing measure is estimated for each subband. This is in agreement with the observations of the excitation spectrum because in voiced sounds, the higher frequency regions usually contain noise-like components that are not modeled appropriately with impulses.

Ther first step in the analysis algorithm is determining the pitch period using an autocorrelation based method. The excitation spectrum is divided into harmonic bands using the pitch period information. Robust pitch detection is required in this step. This is one of the disadvantages of the model. The voiced and unvoiced spectral envelope parameters are estimated using sinusoidal modeling techniques. Voiced/Unvoiced decisions are made for each harmonic subband which is centered around each integer multiple of the f0. Each harmonic subband covers a range that is equal to the f0 value. The pitch period estimate is refined and the parameters are re-estimated for improving robustness in the case of pitch detection errors.

In the synthesis stage, amplitudes, phases and center frequency values are interpolated between frames. Voiced portions of the signal are synthesized using the corresponding sinusoidal model parameters. The unvoiced portions are approximated by bandpass filtered White Gaussian Noise multiplied by the spectral envelope in the frequency domain. Inverse FFT is employed to obtain the unvoiced waveform in time domain. Finally, these voiced and unvoiced portions are summed up to obtain the synthetic segment. The segments are concatenated using weighted overlap-add similar to PSOLA methods. The details of the MBE model can be found in (Griffin, 1987).

The major advantage of the MBE model relies on the fact that the excitation spectrum is represented in detail. It is possible to quantize or interpolate the sinusoidal model parameters. The disadvantages are related to the requirements for robust pitch detection, phase interpolation and voiced/unvoiced harmonic subband decisions. Besides these disadvantages, MBE model has become a popular model for coding and synthesis applications over the years as described in (Hardwick and Lim, 1988), (Nishiguchi *et al.*, 1993), and (Wang *et al.*, 1996).

3.3.3. Glottal Flow Models

Several parametric models exist for estimating and modifying glottal flow waveforms both in the time domain and in the frequency domain. The Liljencrants-Fant (LF) model is a well known example (Fant *et al.*, 1985). It represents the derivative of the glottal flow with the following timing parameters:

- T_0: instant of glottal opening
- T_p and U_0: instant and value of maximum glottal flow
- T_e and E_e: instant and absolute value of the minimum of the glottal flow derivative
- T_a: return phase that can be defined as the absolute time difference between t_e and the projection of the tangent of the glottal flow derivative at t_e
- T_c: instant of glottal closure

In the LF model, the derivative of the glottal flow is composed of two parts. The first part characterizes the glottal flow derivative from the glottal opening to the maximum negative peak. The second segment characterizes the closure of glottis.

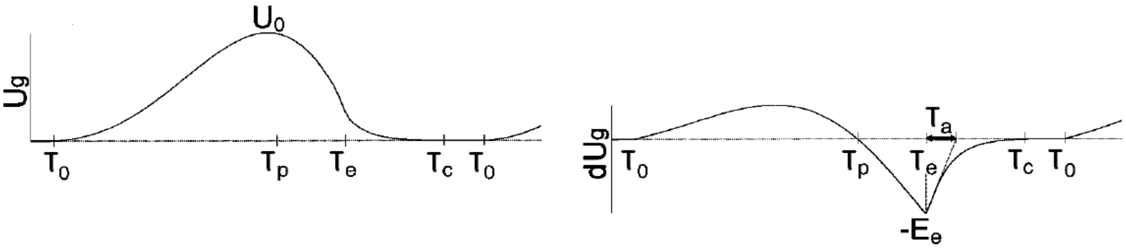


Figure 3.8. Glottal flow (U_g) and its derivative(dU_g) in the LF model (Strik, 1998)

$$\int_0^{T_0} g(t)\, dt = 0 \tag{3.24}$$

$$w_g = \frac{\pi}{T_p} \tag{3.25}$$

$$\varepsilon T_a = 1 - e^{-\varepsilon(T_c - T_e)} \tag{3.26}$$

$$e^{\alpha T_e} sin(\pi w_g T_e) = -1 \tag{3.27}$$

$$E_0 = -\frac{E_e}{e^{\alpha T_e}\sin(w_g T_e)} \tag{3.28}$$

The synthesis parameters E_0 , α, ω_g ,and ε can be derived from the timing parameters using Equations 3.24 to 3.28. Then, the glottal flow g(t) is synthesized using these parameters in Equation 3.29.

$$g(t) = \begin{cases} E_0 e^{\alpha t}\sin(\omega_g t) & ,0 \le t \le T_e \\ \\ -\dfrac{E_e}{\epsilon T_a}[e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_c-T_e)}] & ,T_e \le t \le T_c \le T_0 \end{cases} \tag{3.29}$$

3.4. Modification of the Filter Component

We must estimate the mapping between the source and the target vocal tract parameters in order to modify the source vocal tract parameters to match the target speaker characteristics. The mapping can be obtained in several ways. The primary methods were based on vector quantization and a discrete mapping was estimated. The involvement of continuous mapping methods has led to considerable improvement in voice conversion performance. It is possible to employ a mapping method that inherently estimates continuous mapping functions as in the case of ANNs, GMMs or RBFNs. Another possibility is to employ appropriate interpolation methods in a discrete mapping framework. An example for this approach is codebook mapping which is employed in STASC (Arslan, 1999). As we investigate new methods in the framework described in STASC, we will describe the training and transformation algorithms in the following subsections. These algorithms are modified appropriately in Chapter 5 for the new methods proposed but the general framework is similar.

3.4.1. Training

The training data used in STASC are the same sentences uttered by both source and target speakers. The sentences must be phonetically balanced in order to obtain successful transformation results. The source sentence is aligned first and the target sentence is force-aligned with it next. A Hidden Markov Model (HMM) is generated for each sentence so the alignment procedure is called the *Sentence HMM* method. It is also possible to use manually generated labels. However, the Sentence HMM method enables the automation of the entire training process.

After alignment, we extract the model parameters from the corresponding HMM states in the case of Sentence HMMs. These states are the phonemes in the case of phonetic labeling. The parameters include LSFs for the vocal tract, instantenous f0 values as well as mean and variance of source and target pitch values, durations, and energy values. So, we obtain all parameters in a single training step. Finally, we employ confidence measures to eliminate states that do not possess sufficient similarity in terms spectrum, pitch, duration and energy characteristics. The parameters of all remaining states are kept in the codebooks. The details of the training procedure is shown in Figure 3.9.

Two codebooks are generated for the source and the target speaker separately. The codebook entries include average line spectral frequencies, f0, energy, and duration of each state. The mean and the standard deviation of the f0 values are also included. We use these in the pitch transformation as described in Section 3.6.1.

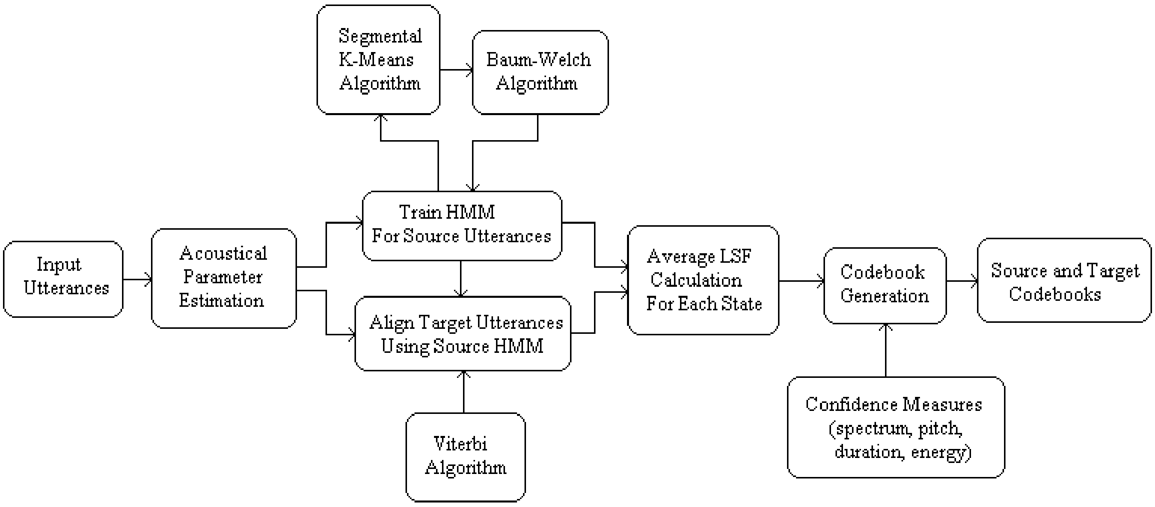

Segmental K-Means Algorithm
Baum-Welch Algorithm
Input Utterances
Acoustical Parameter Estimation
Train HMM For Source Utterances
Align Target Utterances Using Source HMM
Viterbi Algorithm
Average LSF Calculation For Each State
Codebook Generation
Source and Target Codebooks
Confidence Measures (spectrum, pitch, duration, energy)

Figure 3.9. Flow chart for STASC training algorithm

It is also possible to apply pre-filtering using the long-term spectra for the source and the target recordings. The spectrum of the speech frame to be transformed is multiplied with the target long-term spectrum and divided by the source long-term spectrum during FD-PSOLA. This pre-filtering step transforms the long term spectral characteristics of the source speaker to those of the target speaker. This method is useful if the source and target training utterances were recorded in different acoustical environments.

3.4.2. Transformation

At the transformation stage, we analyze the input signal pitch-synchronously. Source/filter decomposition is performed using LP analysis. We modify the source and the filter components separately. We employ FD and TD-PSOLA based modifications on the source signal as described in Section 3.5. In this section, we will focus on the transformation of the vocal tract parameters. Prosodic modifications are described in Sections 3.5 and 3.6.

We use the source LSFs for searching the closest match(es) in terms of vocal tract spectrum in the source codebook. Then, we estimate the output vocal tract spectrum using time varying filtering techniques in the frequency domain. The flowchart of the transformation algorithm is shown in Figure 3.10. Target LSFs are interpolated for vocal tract transformation. For this purpose, we estimate the LP coefficients of the input frame, a_k^s's, and convert them to LSFs, w_k's. We compare these LSFs with the source codebook LSF entries using the distance measure given in Equation 3.30 and determine the entries with minimum distance.

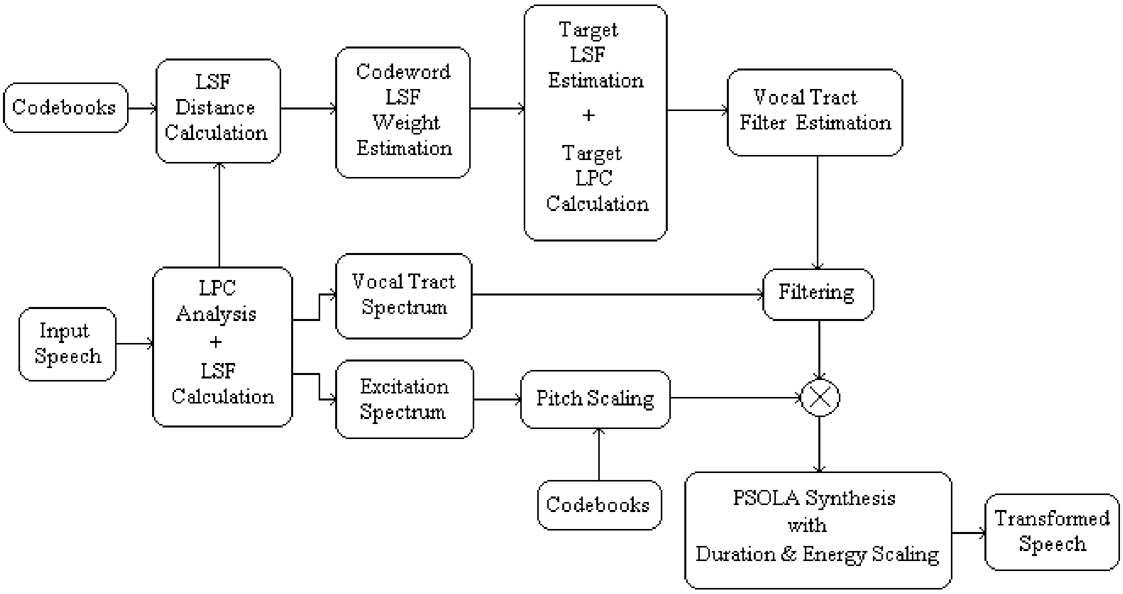

Codebooks
LSF Distance Calculation
Codeword LSF Weight Estimation
Target LSF Estimation + Target LPC Calculation
Vocal Tract Filter Estimation
Input Speech
LPC Analysis + LSF Calculation
Vocal Tract Spectrum
Filtering
Excitation Spectrum
Pitch Scaling
Codebooks
PSOLA Synthesis with Duration & Energy Scaling
Transformed Speech

Figure 3.10. Flow chart for STASC transformation algorithm

$$d^i = \sum_{k=1}^{P} h_k \mid w_k - S_k^i \mid \quad for \quad i = 1, \ldots, L \tag{3.30}$$

where w_k's are the LSFs for the input frame, P is the order of linear prediction analysis, S_k^i is the kth LSF of ith source codebook entry. d^i's correspond to the estimated distances in a codebook of size L. i is the codebook entry index. The LSF weights, h_k's, are estimated using the perceptual weighting criterion given in Equation 3.31. The LSFs with closer values are assigned higher weights because they usually correspond to formant locations.

$$h_k = \frac{1}{argmin(\mid w_k - w_{k-1} \mid, \mid w_k - w_{k+1} \mid)} \quad for \quad k = 1, \ldots, P \tag{3.31}$$

Based on the distances, d^i's, we estimate the normalized codebook weights, v^i's using:

$$v^i = \frac{e^{-\gamma d^i}}{\sum_{l=1}^{L} e^{-\gamma d^l}} \tag{3.32}$$

where γ can be adjusted. Typical values of γ are in the range [0.2, 2] but we have used a value of 1.0 for most of the cases. We estimate the target vocal tract spectrum using these weights and the corresponding target codebook entries:

$$\hat{w}_k^t = \sum_{i=1}^{L} v^i T_k^i \quad for \quad k = 1, \ldots, P \tag{3.33}$$

where $\hat{w}_k^t$ is the k^th^ LSF of the estimated target LSF vector and T^k_i is the k^th^ LSF entry of the i^th^ target codebook entry. We convert the estimated LSFs, $\hat{w}_k^t$'s, to LPC coefficients, $\hat{a}_k^t$'s, and obtain the estimated target vocal tract spectrum $V_t(w)$:

$$V_t(w) = \mid \frac{1}{1 - \sum_{k=1}^{P} \hat{a}_k^t e^{-jkw}} \mid \tag{3.34}$$

The vocal tract filter $H_v(w)$ is estimated as follows:

$$H_v(w) = \frac{V_t(w)}{V_s(w)} \tag{3.35}$$

We multiply the input spectrum with $H_v(w)$ to convert the vocal tract. Note that $V_s(w)$ represents the source vocal tract spectrum. It can be obtained directly from the input speech segment using a_k^s's as in Equation 3.36. As an alternative, we can estimate the source vocal tract spectrum estimated as an approximation to the source codebook LSF entries. In this case, we interpolate the source codebook LSFs using the codebook weights, v_i's, to obtain $\hat{w}_k^s$'s using Equation 3.37. We convert $\hat{w}_k^s$'s to LP coefficients, $\hat{a}_k^s$'s. The source vocal tract spectrum, $V_s(w)$, is calculated using Equation 3.38.

$$V_s(w) = \mid \frac{1}{1 - \sum_{k=1}^{P} a_k^s e^{-jkw}} \mid \tag{3.36}$$

$$\hat{w}_k^s = \sum_{i=1}^{L} v^i S_k^i \quad for \quad k = 1, \ldots, P \tag{3.37}$$

$$V_s(w) = \mid \frac{1}{1 - \sum_{k=1}^{P} \hat{a}_k^s e^{-jkw}} \mid \tag{3.38}$$

3.5. Modification of the Source Component

In many applications, it is necessary to modify the prosodic characteristics of speech signals in a natural and efficient manner. Examples include speech synthesis, voice conversion, speech compression and transcription applications. In concatenative speech synthesis, appropriate segments are collected from a TTS database and concatenated applying necessary modifications in real-time. These modifications include time, pitch and energy scaling, and spectral smoothing to reduce discontinuities across segment boundaries. Efficient time/pitch scale and energy contour modifications are possible employing the methods we describe in this section.. As duration and pitch characteristics serve as important clues in perception of speaker identity, more convincing results are obtained by using prosodic modification methods in voice conversion. As a compression application, the possibility to preserve a time scale compressed version of the speech signal that requires less space for storage can be proposed. On the other hand, a time scale expanded version of a speech signal may be easier to understand for transcription applications. In the next sub-sections, we describe two methods for performing prosodic modifications.

3.5.1. Time Domain Pitch Synchronous Overlap-Add (TD-PSOLA) Algorithm

TD-PSOLA is a simple and effective method for performing prosodic modifications on speech signals. It is well suited for real-time applications. The idea is to process the speech signal on a short-time basis where the segments are obtained pitch synchronously. These segments are concatenated in an appropriate manner to obtain the desired modifications. The overall process is as follows:

- The start and end instants of pitch periods over the voiced regions are determined by pitch marking. The algorithm described in (Gold and Rabiner, 1969) can be used. Pitch detection methods are not suitable for this purpose as the exact instants where the pitch period starts and ends are required.
- Pitch synchronous speech segments are extracted by covering 2 to 5 pitch periods per frame. Windowing is applied.

- Time and pitch scale modifications are performed as described later on and the output is reconstructed using overlap-add synthesis by windowing.

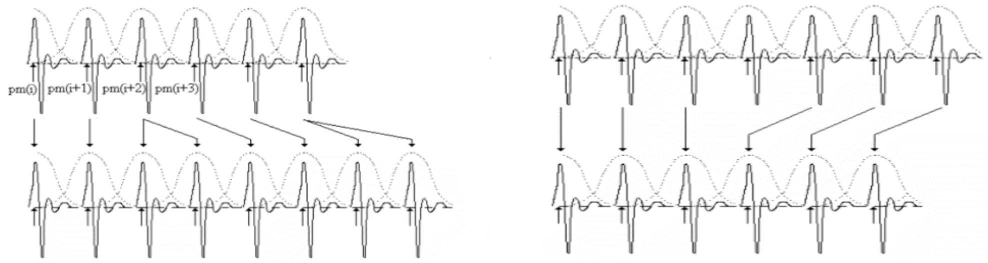


Figure 3.11. Time scale expansion (left) and compression (right)

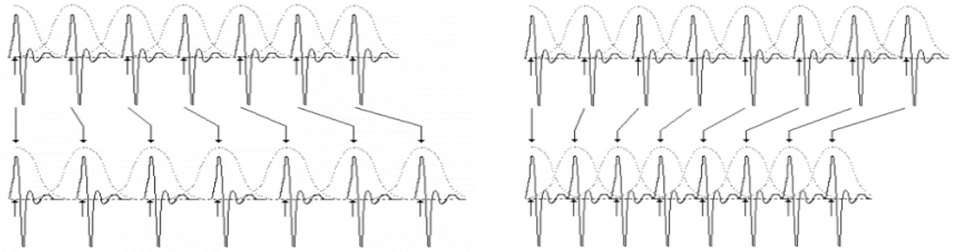


Figure 3.12. Pitch scale compression (left) and expansion (right)

<u>3.5.1.1. Time Scale Modification.</u> It is common practice that when the audio signals are played back at a lower rate, the spectrum of the output is also modified. We know from the properties of the Fourier Transform that expanding the time scale of a signal causes a compression in the frequency domain so the output is a pitch scale compressed version of the original signal. On the other hand, when the time scale is compressed -i.e. when the signal is played back faster, the pitch will be higher. The aim of time scale modification is to prevent these inherent modifications in the signal spectrum while modifying the time axis and obtain an output that has similar spectra as the original signal.

TD-PSOLA modifies the temporal content by repeating or removing integer number of speech segments. Segment repetition produces a signal that is expanded in the time domain while the output using deletion is a time-compressed version of the original signal (Figure 3.11). Repetition/deletion of integer number of frames does not modify the short-time spectral content and distort the relationship between the pitch harmonics.

3.5.1.2. Pitch Scale Modification. In this case, the aim is to modify the short-time spectral content of the signal without modifying its temporal characteristics. The spectral envelope must also remain constant but rather the locations of the pitch harmonics must be modified because modifying the vocal tract will severely effect the perceived speaker identity. As an example, if the spectrum of male speech is shifted to higher frequencies, the output will not only have increased pitch but the quality of a female's or even a child's voice depending on the amount of shift applied. In order to prevent this effect, TD-PSOLA modifies the amount of overlap between successive pitch-synchronous segments as demonstrated in Figure 3.12. It is also clear that pitch scale modification results in the modificiation of the time-scale. Since this is not desired, compensating time-scale modification must be employed.

3.5.1.3. Synthesis By Overlap-Add. In the final step, the output signal is constructed using overlap-add method with windowing. All the procedure described above determines the new locations and overlap ratios of the frames. We multiply each frame with a Hamming window in order to prevent discontinuities. The frames are then concatenated using the overlap ratios obtained in the pitch-scale modification step. The main advantage of TD-PSOLA is its simplicity and efficieny which make it suitable for real-time applications. However, when severe amounts of time and pitch scaling are applied, the output quality degrades. There also other drawbacks:

- The pitch modification introduces scaling of the time axis which must be compensated.
- The duration modification can only be implemented in a quantized manner, with a one pitch period resolution because the time scale factor can have the values 1/2,2/3,3/4,...,4/3,3/2,2/1, etc.
- When performing a duration expansion, the repetition of the segments can introduce metalic artifacts. This can be compansated by reversing even indexed frames during repetition of voiced segments. Unvoiced regions are not modified during both time and pitch scale modification.
- The spectral envelope changes with f0. The effect is apparent for large amounts of pitch scaling. As TD-PSOLA does not modify the spectral envelope, output quality degrades for very large and very small pitch scaling factors.

3.5.2. Frequency Domain Pitch Synchronous Overlap-Add (FD-PSOLA) Algorithm

The second method to be considered for prosodic modifications is FD-PSOLA that operates in the frequency domain. The algorithm is composed of the following steps:

- The short term spectrum of the signal is estimated pitch-synchronously. 2-5 pitch periods are used as the window size. It is possible to use pitch marks as in TD-PSOLA but FD-PSOLA performs considerably well without the pitch marks. However, a robust pitch detection algorithm is reqired.
- The spectral envelope is estimated. Although it is common to employ linear prediction techniques, any spectral envelope estimation method can be used. It is important to obtain a smooth excitation spectrum. As the excitation spectrum is warped to obtain the desired modifications, any region of the excitation spectrum that is not sufficiently flat will be translated to other spectral regions and this may cause distortion.
- To perform pitch-scale modifications different methods can be employed as described in (Moulines and Charpentier, 1990). Two methods are discussed in Sections 3.5.2.1 and 3.5.2.2.
- Time domain segments are overlap-added as described in Section 3.5.2.3.

3.5.2.1. Harmonics Elimination-Repetition. In this method, we first determine the pitch harmonics. The harmonics are then eliminated (repeated) for pitch scale expansion (compression). However, this method has its own problems. Firts, precise estimation of the pitch value is required to determine the pitch harmonics. Next, the phase coherence between the harmonics of the spectrum must be maintained in order to ensure good quality. For this purpose, the original frames should be set at locations corresponding to maximal source excitation which occurs at the glottal closure instants.

3.5.2.2. Spectral Compression-Expansion. The original freqency axis is linearly warped using the pitch scale factor β. The synthetic DFT coefficients are estimated by Equation 3.39 where k_v is obtained by truncating the real value k/β. Note that the original DFT coefficient index is denoted by k. The weight α is found by Equation 3.40. As the local properties of the spectrum are translated from one spectral region to another, this method may cause distortion. However, if the spectral envelope is tracked well, the distortion will be less than TD-PSOLA or harmonics elimination-repetition for pitch scale expansion.

$$Y(k_s) = (1-\alpha)X(k_v) + \alpha X(k_v+1) \tag{3.39}$$

$$\alpha = k_s - \frac{k}{\beta} \tag{3.40}$$

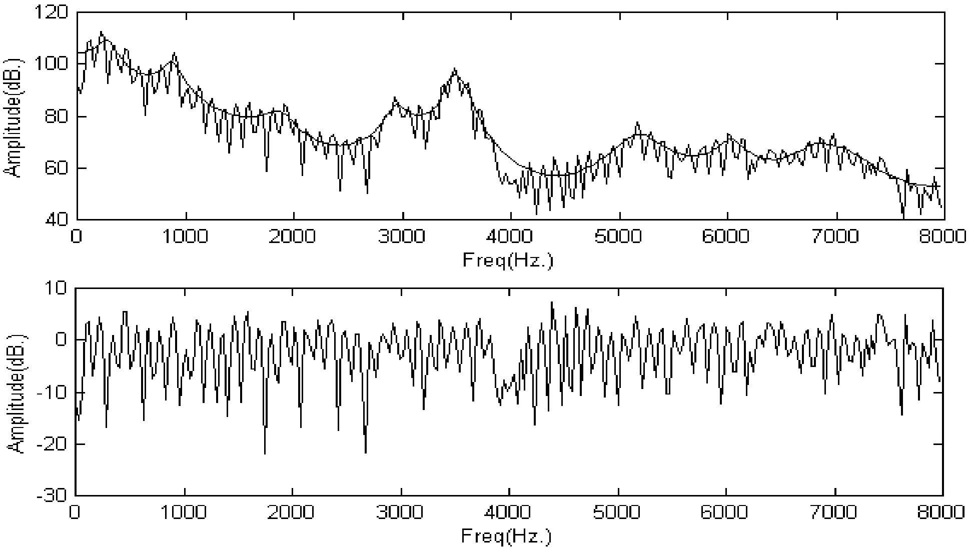


Figure 3.13. Original FFT and LPC spectrum (top), and excitation spectrum (bottom)

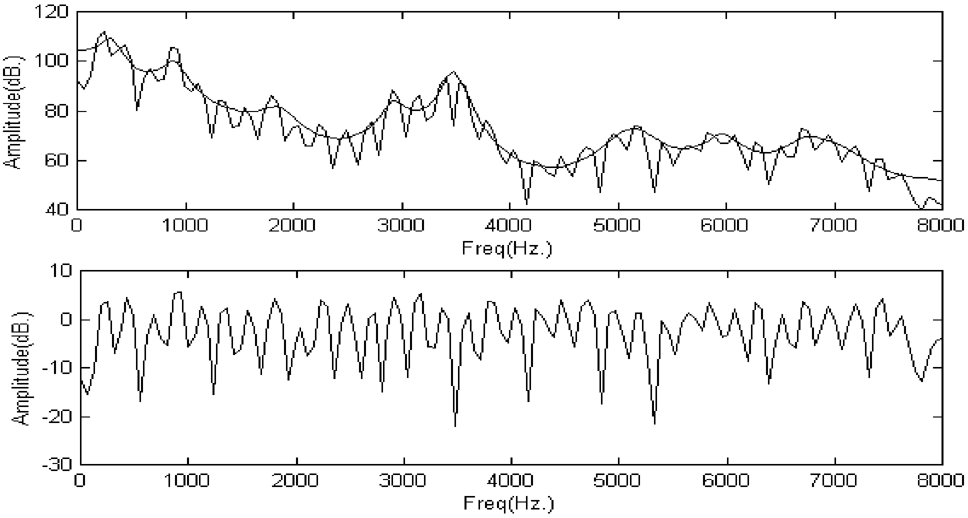


Figure 3.14. Modified FFT spectrum (top) and modified excitation spectrum (bottom) for a pitch scaling ratio of 2.0 using compression/expansion technique

When decreasing the pitch, an empty region appears at the high frequencies. In order to generate an acceptable spectral distribution, either the lower part of the spectrum is copied or the higher part is folded to the empty region. Figures 3.13 and 3.14 demonstrate the method of spectral expansion for increasing the pitch.

3.5.2.3. Synthesis. The Short Time Fourier Transform (STFT) plays a key role in the analysis, modification and synthesis stages of the FD-PSOLA algorithm (Moulines and Verhelst, 1995). At the analysis stage, the speech samples x(n) are windowed by the analysis window $h_u(n)$ pitch-synchronously. If we assume that the windowing function $h_u(n)$ is centered around time t=0, is of finite duration T_u, is symmetric and is the impulse response of a lowpass filter, the analysis short-time signal $x(t_a(u),n)$ associated to the analysis time instant $t_a(u)$ can be represented as in Equation 3.41. Its DFT is given by Equation 3.42.

$$x(t_a(u), n) = h_u(n)x(t_a(u) + n) \tag{3.41}$$

$$X(t_a(u), w) = \sum_{-\infty}^{+\infty} h_u(n)x(t_a(n) + n)e^{-jwn} \tag{3.42}$$

The STFT $X(t_a(u),w)$ is modified as described above to produce the synthesis spectrum $Y(t_s(u),w)$. The problem is that the modified spectra may no longer be a valid STFT in the sense that a signal which has $Y(t_s(u),w)$ as its STFT may not exist. In this case, we wish to obtain a signal y(n) which has its STFT as close as possible to the desired synthesis spectra. If we denote the STFT of y(n) as an approximation to the desired signal spectrum:

$$\hat{Y}(t_s(u), w) = \sum_m f_u(m)y(t_s(u) + m)e^{-jwm} \tag{3.43}$$

The problem of estimating y(n) can be solved by least-squares fitting to minimize the following summation:

$$\sum_u \int_{-\pi}^{+\pi} \mid Y(t_s(u), w) - \hat{Y}(t_s(u), w) \mid^2 \, dw \tag{3.44}$$

The sum is over all time instants $t_s(u)$ where $Y(t_s(u),w)$ is defined. $f_u(n)$ is the synthesis window. The solution to the least squares problem is given by:

$$y(n) = \frac{\sum_u y_w(u, n - t_s(u)) f_u(n - t_s(u))}{\sum_u f_u^2(n - t_s(u))} \tag{3.45}$$

$$y_w(u, n) = \frac{1}{2\pi} \int_{-\pi}^{+\pi} Y(t_s(u), w) e^{jwn}\, dw \tag{3.46}$$

The synthesis algorithm is similar to weighted overlap-add. The successive short-time synthesis signals are combined with appropriate weights and time-shifts. If we choose the synthesis window such that $\Sigma_u f_u^2(n-t_s(u)) = 1$, the synthesis operation is simplified. Let's consider an example to see how the synthesis algorithm operates:

- Let $h_u(n) = h(n)$ be the fixed analysis window of length L for a constant analysis rate $t_a(u) = uR$ where R<L (so the analysis frames are overlapping)
- To compute x(uR,n), the signal is advanced R points in time and windowed to obtain x(uR,n) = h(n)x(n+uR). The STFT is calculated to obtain X(uR,w)
- Let's have no modification in order to demonstrate the OLA synthesis method. The inverse STFT is computed to recover the windowed segments x(uR,n). The result is windowed again with the synthesis window $f_u(n)=h(n)$ to obtain h(n)x(uR,n). Of course this is not the case when considering pitch-scale modifications, so the synthesis window will be different than the analysis window. As all these segments were positioned around time origin during the analysis, they have to be delayed to move each one back to its original location along the time axis (i.e. around time uR for segment number u). The result is given by a time-varying normalization weight as follows:

$$Y(n) = \frac{\sum_u h(n - uR) x(uR, n)}{\sum_u h^2(n - uR)} = \frac{\sum_u h^2(n - uR) x(n)}{\sum_u h^2(n - uR)} = x(n) \tag{3.47}$$

Thus, OLA synthesis formula reconstructs the original signal if the analysis STFT $X(t_a(u),w)$ is a valid STFT. This is the case when we have no modifications. If modifications are performed, Equation 3.47 will estimate the signal which has a maximally close STFT to $X(t_a(u),w)$ in the least-squares sense. In fact, this procedure provides a framework suitable for different approaches. Reconstruction of speech signals using only the short time magnitude spectra, OLA time-scaling, synchronized OLA time-scaling, and WSOLA are examples.

The main drawback of FD-PSOLA is the requirement for high computational power. A comparative study of several PSOLA methods can be found in (Moulines and Charpentier, 1990) and (Violaro and Böeffard, 1998).

In STASC, the excitation spectrum is modified using the compression/expansion technique of FD-PSOLA for pitch scale expansion. However, for pitch scale compression we have used TD-PSOLA to reduce distortion by avoiding the empty spectral region in lowering the pitch. In this case, the excitation spectrum is not modified but the synthesis time instants are arranged in order to obtain the desired amount pitch scale compression. It is also possible to modify the excitation magnitude spectrum employing the method described in Section 3.5.3. After all these modifications, we multiply the excitation spectrum with the transformed vocal tract spectrum and obtain the synthesis segment using Equation 3.47. Duration and energy scaling can be performed at this stage also as shown in Figure 3.10.

3.5.3. Excitation Transformation

Although the spectral envelope contains important clues of speaker identity, the residual signal possesses the rest of the information that may be useful for both speaker-specific modeling and modification. In the following paragraphs, we describe the excitation spectrum modification method as employed in (Arslan, 1999). In this approach, the magnitude of the excitation spectrum is transformed to match the target speaker characteristics.

In STASC, an excitation transformation filter is constructed using the selected codebook entries for vocal tract conversion. The same set of weights are used as in Equation 3.48 to construct the filter. Using this filter and the converted vocal tract spectrum, the speech spectrum can be obtained as in Equation 3.49.

$$H_g(w) = \sum_{i=1}^{L} v_i \frac{U_i^t(w)}{U_i^s(w)} \tag{3.48}$$

$$Y(w) = H_g(w) H_v(w) X(w) \tag{3.49}$$

where X(w) is the input speech spectrum. The frames are then overlap-added pitch synchronously to generate the output.

It is also possible to modify the phase component. In (Kain, 2001), the author describes a method for transforming both the magnitude and the phase component of the excitation spectrum. The excitation magnitude spectrum vectors are employed to train a GMM to account for different vectors that can be observed in different phonemes. The phase vector of the centroid of each class is used in synthesizing the output.

3.5.4. Other Methods

The sinusoidal parameters are used in several studies for modification of speech prosody (Quatieri and McAulay, 1992). As these parameters are a set of amplitudes, frequencies and phases, it is pretty straightforward to realize pitch, time-scale and energy modifications. Phase vocoding techniques are also used for modifying speech prosody. The baseline phase vocoder system is described in (Flanagan and Golden, 1966). Pitch scaling can be obtained by applying appropriate amounts of time-scaling and resampling (Laroche and Dolson, 1999). In (Tang *et al.*, 2001), pitch, duration, and energy modification are performed using a phase vocoder without explicit pitch detection.

3.6. Modeling and Transforming the Pitch

Pitch contour modeling has been addressed in many studies on speech synthesis. For improving the naturalness of the synthesizer output, appropriate intonational patterns should be generated automatically. Particularly the study in (Taylor, 1992) describes many approaches in pitch contour modeling. Here, we will only cover several techniques for demonstrating the approaches that can be used. In Section 5.4, we describe a new pitch contour modeling and transformation method, and compare it with the mean/variance model.

3.6.1. Mean/Variance Model

The simplest approach for modeling the pitch contours is to assume that the f0 values come from a single Gaussian distribution and estimate the mean and the variance of the distribution. For this purpose, after the pitch values are detected, only voiced segments are used for estimating the parameters of the distribution. Once the mean and the variances of two speakers are estimated, transformation becomes easy. Let's define two random variables S and T from two separate Gaussian distributions with means and variances ($\mu_{s,}$ $\sigma^2{}_s$) and (μ_t, $\sigma^2{}_t$). Let's denote the pdf's of these distributions as $f_S(s)$ and $f_T(t)$ respectively. Assuming a linear transformation rule from source to target pitch, the target pitch value is given by:

$$t = as + b \tag{3.50}$$

The problem is to estimate a and b using the source and the target pitch statistics. The pdf of a linear transformation of a random variable is given by:

$$f_T(t) = \mid \frac{\partial g^{-1}(t)}{\partial t} \mid f_S(g^{-1}(t)) \tag{3.51}$$

$$t = g(s) = as + b \ \Rightarrow g^{-1}(t) = \frac{t-b}{a} \ \Rightarrow \ \frac{\partial g^{-1}(t)}{\partial t} = \frac{1}{a} \tag{3.52}$$

If we assume that a>0:

$$f_T(t) = \frac{1}{a} f_S(\frac{t-b}{a}) \tag{3.53}$$

As both distributions are approximated by two Gaussian distributions with means and variances ($\mu_{s,}$ $\sigma^2{}_s$) and (μ_t, $\sigma^2{}_t$), we have:

$$\frac{1}{\sqrt{2\pi\sigma_t^2}} exp(-\frac{(t-\mu_t)^2}{2\sigma_t^2}) = \frac{1}{a} \frac{1}{\sqrt{2\pi\sigma_s^2}} exp(-\frac{(s-\mu_s)^2}{2\sigma_s^2}) \tag{3.54}$$

Taking the logarithm of both sides of the equation and replacing t = as+b, we get:

$$-\frac{(as+b-\mu_t)^2}{2\sigma_t^2} = -\frac{(s-\mu_s)^2}{2\sigma_s^2} + \log(\frac{\sigma_t}{a\sigma_s}) \tag{3.55}$$

Considering s^2 terms:

$$-\frac{a^2}{2\sigma_t^2} = -\frac{1}{2\sigma_s^2} \Rightarrow a = \frac{\sigma_t}{\sigma_s} \Rightarrow \log(\frac{\sigma_t}{a\sigma_s}) = 0 \tag{3.56}$$

The log term is thus eliminated. Using this value for a in Equation 3.55, and taking square root of both sides:

$$\frac{-(\frac{\sigma_t s}{\sigma_s} + b - \mu_t)}{\sigma_t} = \frac{-(s-\mu_s)}{\sigma_s} \Rightarrow b = \mu_t - \frac{\sigma_t \mu_s}{\sigma_s} \tag{3.57}$$

In fact, this pitch transformation method is used in several studies and performs quite well as the mean and variance of the pitch values are successfully converted to match the target speaker characteristics. This model is used in STASC for modeling and transforming pitch characteristics. For this purpose, a time-varying pitch scale factor β(t) is estimated using the instantenous pitch value on the signal to be transformed. We estimate β(t) using the source and the target pitch statistics as in Equation 3.58.

$$\beta(t) = \frac{af_0^s(t)+b}{f_0^s(t)} \quad where\ a = \frac{\sigma_t}{\sigma_s}, \quad and\ b = \mu_t - \mu_s\frac{\sigma_t}{\sigma_s} \tag{3.58}$$

where $f_0^s(t)$ is the instantenous source f_0 value, μ_s and μ_t are source and target mean f_0 values, σ_s and σ_t are the standard deviations of the source and target f_0's respectively.

3.6.2. Sentence Codebooks

It is possible to generate a pitch contour codebook for modeling a speaker's pitch characteristics on the sentence level as described in (Chappel and Hansen, 1998). In this case, the advantage is to be able to use real pitch contours in synthesis or modification. However, the codebook should contain sufficient number of pitch contours. This is impossible when the aim is to be able to synthesize all kinds of pitch contours. This approach works well for a limited vocabulary or for specific applications in which the variability of the pitch contours are restricted.

3.6.3. Fujisaki's Model

In this model, pitch (or intonation) contour's are assumed to be composed of two different components: the phrase and the accent. The production of these contours is modeled as a filtering action of the glottal oscillation mechanism. Impulses and step functions are input to two critically damped filters to generate the corresponding intonation contour. The first filter accounts for the phrase component and the second for the accent component. Phrase components are generated by using impulses as input to the phrase filter and the accent components are obtained by using step functions to excite the accent filter (Fujisaki and Kawai, 1982).

$$\begin{aligned}\ln F_0(t) = \ln F_{min} + \sum_{i=1}^{I} A_{p_i} G_{p_i}(t - T_{0_i}) + \\ \sum_{j=1}^{J} A_{a_j}(G_{a_j}(t - T_{1_j}) - G_{a_j}(t - T_{2_j}))\end{aligned} \tag{3.59}$$

$$G_{p_i}(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t) \ , \ t \geq 0 \\ 0 \ , \ t < 0 \end{cases} \tag{3.60}$$

$$G_{a_j}(t) = \begin{cases} \min(1 - (1 + \beta_j t)\exp(-\beta_j t), \theta) \ , \ t \geq 0 \\ 0 \ , \ t < 0 \end{cases} \tag{3.61}$$

Here, G_{pi}'s are the filters corresponding to phrase commands (Equation 3.60) and G_{aj}'s are the filters corresponding to the accent commands (Equation 3.61). Successive phrases can be added to the tails of the previous ones creating the contours using Equation 3.59. The accent commands are useful for controlling the duration of each movement in the intonation contour. By the use of step functions of sufficient duration, the accent commands can be shaped. In (Narusawa *et al.*, 2002), the authors describe methods for estimating the model parameters automatically.

3.7. Subband Processing and DWT

In this section, we summarize both the motivation and the theoretical background for subband processing. We start with a brief description of the human auditory system. Next, we summarize the Discrete Wavelet Transform (DWT).

Human auditory system serves as a frequency analyzer. This function is performed by the basilar membrane which resides in the inner ear (cochlea). The principal structures of the human auditory system is shown in Figure 3.15. When tones of different frequencies are transmitted to the inner ear, different regions in the basilar membrane vibrate enabling analysis of the frequency content of the incoming signals. This behaviour is modeled by a bank of bandpass filters in auditory scene analysis, coding and recognition applications.

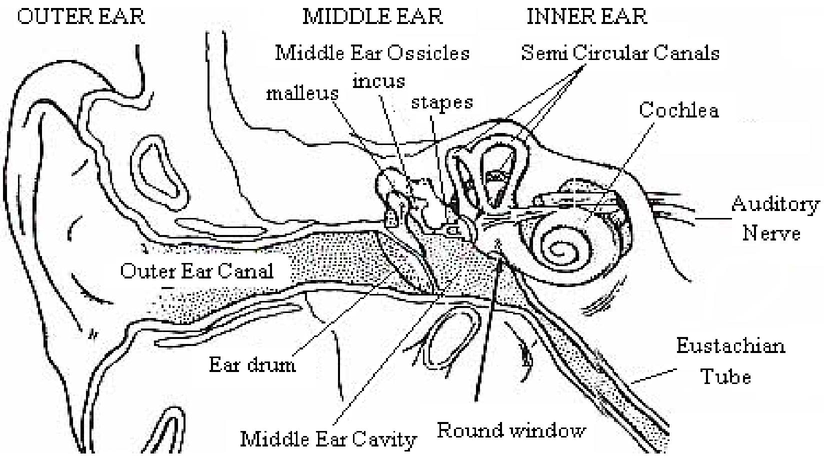


Figure 3.15. Human auditory system (Kenny, 2001)

In this study, we investigate the use of subband processing in voice conversion, particularly in vocal tract conversion. Chapter 5 describes two new methods based on subband processing for this purpose. Multi-resolution signal processing techniques serve as a useful tool for subband processing. Wavelet Transform provides a practical framework.

Continuous Wavelet Transform (CWT) of a finite-energy signal f(t) is defined by the relation in Equation 3.62 where $\varphi(\frac{t-b}{a})$ is a time-shifted and time-scaled copy of the mother wavelet $\varphi(t)$. Note that $\varphi(t)$ is a fixed function in $L^1(\mathbb{R}) \bigcap L^2(\mathbb{R})$ as described in (Torresani, 1999).

$$T_f(b, a) = \frac{1}{a} \int f(t)\varphi(\frac{t-b}{a})\, dt \tag{3.62}$$

The discrete version of CWT, called the Discrete Wavelet Transform (DWT) is obtained using digital filtering and subsampling operations as given in Equation 3.63. These equations correspond to a single step of decomposition for DWT. Subband decomposition at the desired scale can be performed by a cascade of these operations.

$$\begin{aligned} y_{high}[k] &= \sum_n x[n]g[2k-n] \\ y_{low}[k] &= \sum_n x[n]h[2k-n] \end{aligned} \tag{3.63}$$

Equation 3.64 corresponds to Inverse Discrete Wavelet Transform (IDWT) for reconstructing the signal from its subband components $y_{low}[k]$ and $y_{high}[k]$. The filters h[n] and g[n] are not independent of each other but their filter coefficients has the relation given in Equation 3.65. These are known as Quadrature Mirror Filters (QMF). If the filter pairs h[n] and g[n] form an orthonormal pair, then IDWT will produce the input signal x[n] exactly. This is known as perfect reconstruction (PR). The orthonormal filter pairs developed by Daubechies are well known and used in many applications.

$$\hat{x}[n] = \sum_{k=-\infty}^{\infty} (y_{high}[k]g[-n+2k]) + (y_{low}[k]h[-n+2k]) \tag{3.64}$$

$$g[L-1-n] = (-1)^n h[n] \tag{3.65}$$

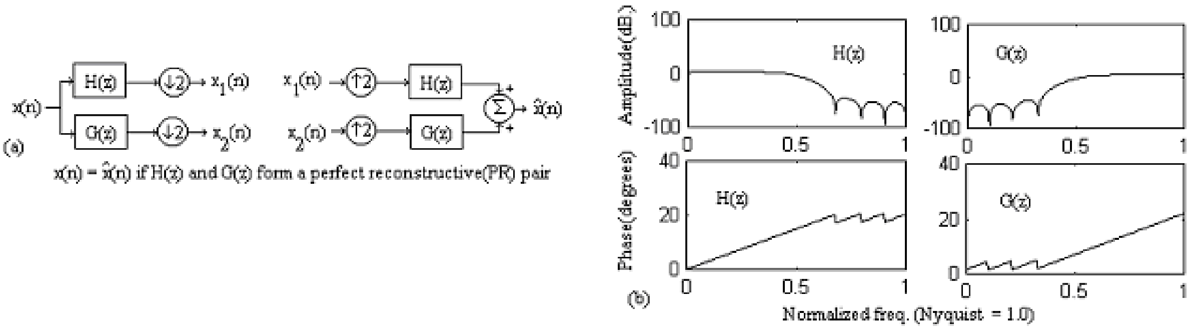


Figure 3.16. DWT flowchart for one level of decomposition and reconstruction (left), the magnitude and phase responses of the filter pair (right)

DWT can be implemented either using a pyramid structure or using a lattice filtering approach. In this study, the pyramid structure shown in Figure 3.16 is implemented. One can use a cascade of this basic structure to implement DWT decomposition and reconstruction filterbanks of any order (Burrus *et al.*, 1998). DWT based subband decomposition is employed in training for voice conversion as discussed in Section 5.1. The transformation stage involves both DWT based subband decomposition on the input signal and DWT based subband reconstruction to obtain the voice conversion output. Following characteristics of the DWT make it an attractive tool for designing filterbanks:

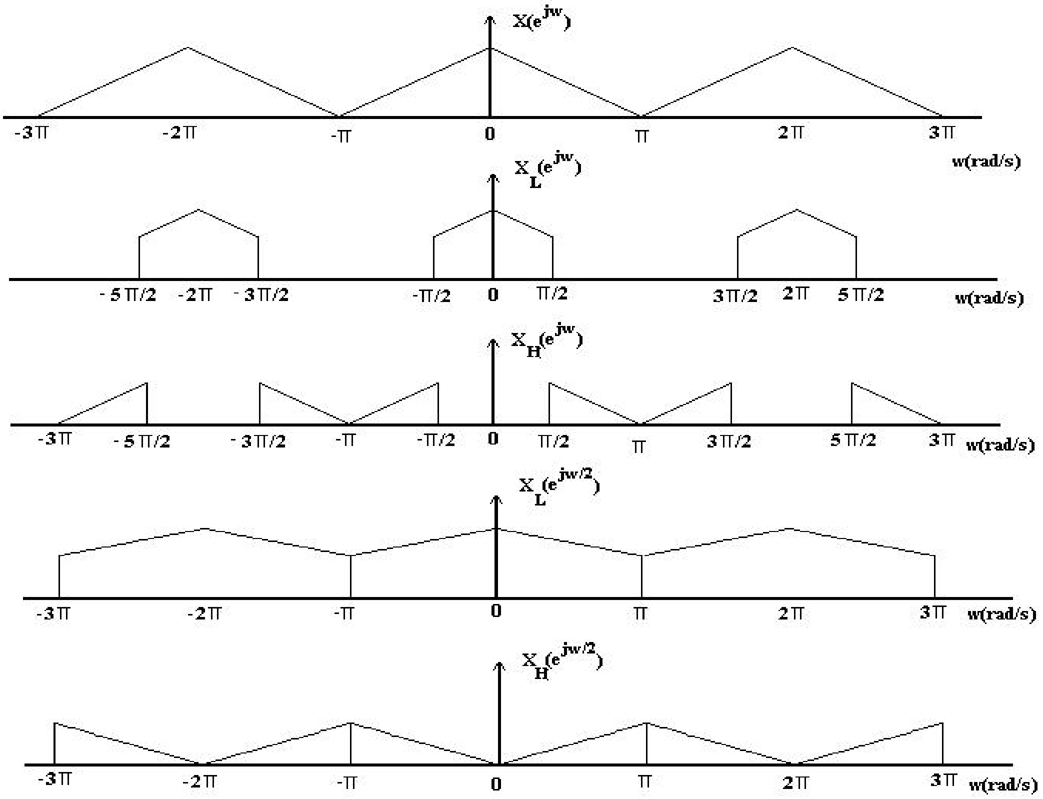


Figure 3.17. Lowpass & highpass filtering followed by decimation

- Perfect reconstruction is guaranteed if appropriate filter pairs are used.
- FIR filters can be used which are guaranteed to be stable having linear phase and by increasing the order of the filters, a filterbank with sharp cut-off filters can be designed. This is useful for bandpass filtering.
- Subband decomposition (reconstruction) can be realized fully in time-domain using convolution and decimation (interpolation).

- Aliasing can be prevented since appropriate lowpass and highpass filters are used before decimation and after interpolation.

The baseband signal corresponding to the higher subband have inverted spectrum due to decimation in the case of one-level decomposition. The reason is that decimation in time domain corresponds to stretching the spectrum as shown in the Figure 3.17. After each high-pass filter, spectral inversion due to shifting is observed. In Figure 3.17, $X(e^{jw})$ is the original signal spectrum. The signal is first lowpass and highpass filtered with filters having cut-off frequencies at π/2 rad/s. $X_L(e^{jw})$ and $X_H(e^{jw})$ denotes the filtered versions of the original spectrum. Then, we apply decimation by 2 in time and obtain the baseband signal spectra $X_L(e^{jw/2})$ and $X_H(e^{jw/2})$.

By comparing $X(e^{jw})$ with $X_H(e^{jw/2})$, we observe that the spectrum is inverted. We must consider this inversion when the purpose is to bandpass filter a signal using DWT. As an example we can consider a DWT filterbank with order 3. This filterbank will produce $2^3 = 8$ baseband signals. Using these baseband signals with sequences of zeros in the appropriate channels of the reconstruction filterbank, we can obtain the bandpass filtered versions of the original signal.

4. SUBJECTIVE TESTS FOR PERCEPTION OF SPEAKER IDENTITY

4.1. Subjective Assessment of Frequency Bands for Perception of Speaker Identity

In this part, we have designed subjective tests to assess the importance of frequency bands for perception of speaker identity. Section 4.1.1 starts with a brief description of the procedure and describes the filterbank used for subband decomposition for the construction of the test database. In Section 4.1.2, we focus on the design of the subjective tests, the content of the database used and evaluation methods. Section 4.1.3 presents the results obtained. In (Ormancı *et al.*, 2002), we have only described the first two parts of the subjective tests performed in this chapter (Sections 4.1.2.1 and 4.1.2.2). Part III, as described in Section 4.1.2.3 is a complementary work which provides further explanations on the evidence.

4.1.1. General Framework

The procedure can be briefly described as a test in which the subjects were required to listen to several utterances along with the bandpass filtered versions of these utterances and provide a subjective score on the similarity of speaker identities. The bandpass filterbank used in the experiment has been adapted from MPEG coding (ISO/IEC, 1993) by taking the frequency bands relevant to speech. Figure 4.1 and Table 4.1 show the cut-off and center frequencies and the magnitude responses of the filters employed in the filterbank. 10 FIR filters of order 50 were used in the design.

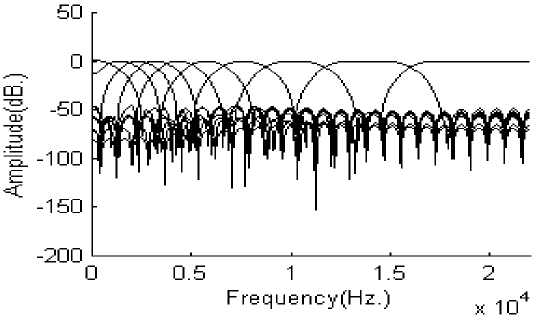
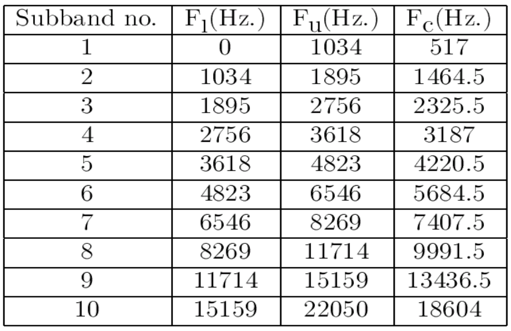


Figure 4.1.Magnitude responses of the bandpass filters



Subband no.	F_l(Hz.)	F_u(Hz.)	F_c(Hz.)
1	0	1034	517
2	1034	1895	1464.5
3	1895	2756	2325.5
4	2756	3618	3187
5	3618	4823	4220.5
6	4823	6546	5684.5
7	6546	8269	7407.5
8	8269	11714	9991.5
9	11714	15159	13436.5
10	15159	22050	18604

Table 4.1.Upper (F_U), lower (F_L) cutoff, and center frequencies of the bandpass filters

4.1.2. Methodology

We have used 10 words in Turkish uttered by eight native Turkish speakers in different ages (four female, four male) to generate the database. The words cover all the phonemes of the Turkish language. The primary advantage of a word database is that the intonational differences are reduced. So, the subjects are less likely to focus on the prosodic characteristics of the speakers when identifying them. Using a sentence database without normalized prosody will always lead to more biases related to prosody than a word database recorded in the same conditions. This is due to the fact that speakers are more likely to utter words in a smooth manner without employing intonational patterns much.

First, each word in the database is decomposed using the bandpass filters with properties described in Figure 4.1 and Table 4.1. Three separate lists of speech files were prepared. In the first list, the original and bandpass filtered version of the same word is used. Different words were used to generate the pairs in the second list in order to reduce the dependence of the results on prosody further. This was the case because the database still possesses prosodic differences in the word level. In the last list, we use the f0 normalized versions of the words and the pairs consisted of different word pairs as in the second list. By this method we ensure complete removal of the intonational information in the database. The order of the speech file pairs were arranged randomly in all lists in order not to have pairs containing the same word one after another. A graphical user interface in VOX (See Appendix A) was designed by which the subjects were able to listen to the speech files and score them accordingly.

Score (% similarity)	Idea reflecting the score
100	I think the speakers are the same
75	They sound like the same but not for sure
50	I have no idea
25	They sound like different but not for sure
0	The speakers are different

Table 4.2. Scoring instructions

Are speakers the same?	Score	Modified score
Yes	100	100
Yes	75	75
Yes	50	50
Yes	25	25
Yes	0	0
No	100	0
No	75	25
No	50	50
No	25	75
No	0	100

Table 4.3. Modification of scores

In order to get a measure of the consistency of the subjects, ten pairs in the list are arranged such that the original versions and the bandpass filtered versions were uttered by different speakers. The subjects were expected to recognize that the word in the corresponding pair is uttered by different speakers for consistency. The consistency information is used in the calculation of the scores for the second part of the experiment. We had 20 subjects for Part I and Part II of the test. In Part III, 10 subjects have provided the scores. In all parts, they were asked to rate the probability of each pair of words being uttered by the same speaker using the scoring instructions given in Table 4.2.

4.1.2.1. Part I. We use the first list in this part. Each subject listened to the files in the list and provided scores as described in Table 4.2. The scoring method used was summation of the modified scores as given in Equation 4.1. This score calculation method does not rely on the consistency of the subjects. We used Table 4.3 to modify the scores. Modification is required because some of the pairs were uttered by different speakers. As an example, consider the case when the speakers in the utterance pairs were different. In this case, a score of "0" for the similarity of the speakers should be regarded as the perfect answer and assigned a score of "100" instead of "0". Note that similarity_n denotes the scores in Table 4.2 and $M[\text{similarity}_n]$ denotes the modification performed as given in Table 4.3. The means and standard deviations of the scores provided for each subband are then calculated and used as a measure of the importance of a subband in perceiving a speaker's identity (Table 4.5).

$$score_{i,j} = \sum_{n=1}^{N} M[similarity_n], \ i = 1, \ldots, \# \ of subjects \qquad (4.1)$$
$$N = \# \ of \ sound \ file \ pairs \ in \ subband \ j$$

4.1.2.2. Part II. The second part of the experiment is similar to the first part with two major differences:

- Each pair contained different words as the original utterance and the bandpass filtered utterance. We have observed that the subjects were likely to use prosodic clues in their decisions in Part I. We have used different words in each pair to prevent the use of these clues.
- We have calculated consistency scores for each subject to be used in obtaining the statistics in order to reject inconsistent responses.

We employ Table 4.4 to obtain the consistency score for each subject. The consistency score of each subject is determined by summing up the consistency score assigned for each pair of sound files. However, if the consistency score is below a threshold, the responses are not included in the subband statistics. The least consistent 10 percent of the subjects were rejected. We have used Equation 4.1 to obtain the subband scores. The means and the standard deviations of the scores for each subband is given in Table 4.5.

Are speakers the same?	Score	Consistency score
Yes	100	1.00
Yes	75	0.75
Yes	50	0.00
Yes	25	-0.75
Yes	0	-1.00
No	100	-1.00
No	75	-0.75
No	50	0.00
No	25	0.75
No	0	1.00

Table 4.4. Consistency score for each decision

4.1.2.3. Part III. In this part, all the recordings are re-synthesized at fixed f0 to remove the intonational bias on the test results. The results of the first two parts indicate that the second subband was the most important frequency band in perception of speaker identity as demonstrated in Section 4.1.3. Although we have employed a method to reduce the prosodic biases in Part II, we were not able to investigate the effect of f0 on the responses of the subjects. Fixed f0 resynthesis was performed using FD-PSOLA as described in Section 3.5.2. We provide a fixed f0 value and perform appropriate amount of pitch scaling. Figure 4.2 shows the output pitch contour after f0 normalization. Next, we decompose the signals into subbands using the filterbank described above.

We have extended the number of pairs in the list to 90. 10 pairs contain different speakers for scoring the consistency of the subjects as in Part II. The remaining 80 pairs contain 10 words from each speaker with one of the 10 possible subband versions. As we have eight speakers, the number of all pairs sum up to 80. We have prepared two lists of 90 pairs that were f0-normalized at 100 Hz and 180 Hz. Using two different target f0s, we are able to investigate the effect of the specific f0 value in the test results.

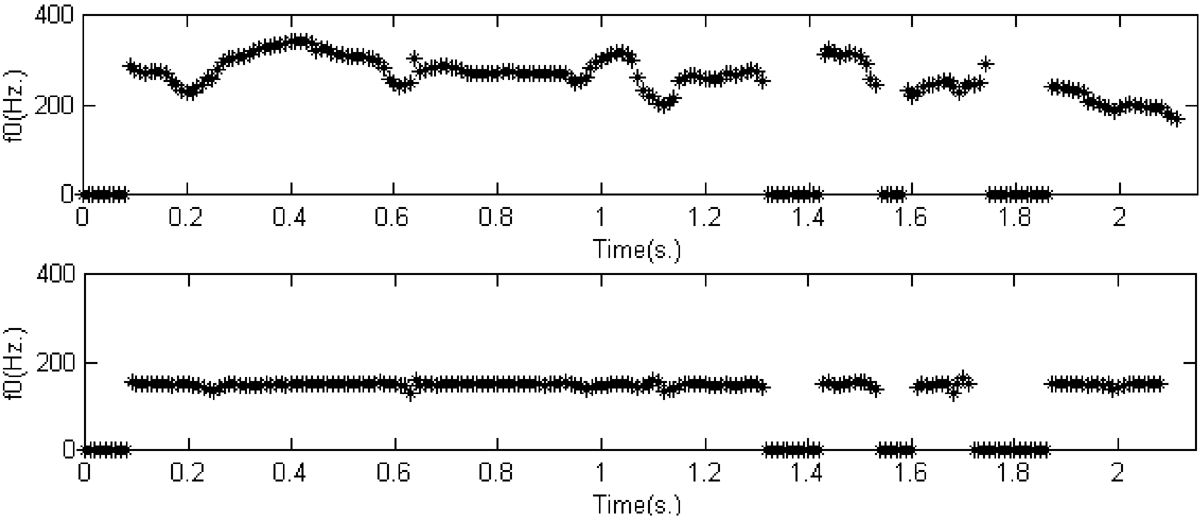


Figure 4.2. Intonation normalization using FD-PSOLA. Pitch contour of a female speaker for the sentence “Kaza nedeniyle ulaşım aksadı” (top) and pitch contour after f0 normalization at 150 Hz (bottom)

4.1.3. Results

The means and the standard deviations of the scores assigned for each subband are shown in Table 4.5. The second subband located between [1034 Hz, 1895 Hz] was assigned the highest mean score for all three parts. The standard deviation was also lower for this subband. The subband of second importance turned out to be the third one which covers the frequency range [1895 Hz, 2756 Hz] with a center frequency of 2325.5 Hz.

If we compare the scores provided in Part I and II, we observe that the scores for Part I are higher because of the prosodic information present in the recordings used. With the prosody information available, the subjects were more accurate in their decisions. We have also calculated the consistency scores for Part I using Table 4.4 in order to compare them with the scores of Part II. The mean consistency score for Part I turned out to be 27.39 which is considerably higher than 19.03 - the mean consistency score for Part II. Note that these values were in the range [8.00,30.25]. So, the subjects had more information for the perception of speaker identity in Part I in which more prosodic clues were present. In Part III, the mean consistency score for a fixed f0 of 100 Hz was 16.00 (in the range [12.00, 20.00]). For the case of 180 Hz, the mean consistency score was 17.00 (in the range [7.00, 27.00]). Note that in the f0-normalized case, the consistency scores were lower than Part I, and Part II as the prosodic clues were removed further by f0 normalization. We also observe higher scores for the higher frequency subbands when f0 was fixed at 180 Hz as compared to 100 Hz. In the 100 Hz. case, lower subbands were assigned higher scores.

	Part I		Part II		Part III(f0=100 Hz.)		Part III(f0=180 Hz.)	
Subband no.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
1	82.1	29.2	63.8	37.5	68.1	34.1	66.7	33.2
2	86.3	25.6	85.0	36.0	80.6	32.7	76.4	32.6
3	84.6	29.6	73.4	32.4	73.6	29.0	58.3	34.3
4	63.2	34.5	64.7	36.0	65.3	43.8	63.9	35.6
5	62.1	37.5	65.2	35.0	48.6	41.5	58.3	39.3
6	77.7	32.0	69.7	24.8	44.4	41.6	59.7	43.8
7	73.7	36.7	67.2	36.2	51.4	44.9	70.8	40.4
8	68.6	34.5	62.2	39.0	73.6	29.0	75.0	34.3
9	76.4	32.4	65.3	37.4	50.0	43.7	41.7	42.0
10	59.3	29.5	57.5	37.2	16.7	38.3	30.6	45.8

Table 4.5. Means and standard deviations of the scores assigned to each subband for all parts

We have also calculated the consistency scores for the first and the last 25 pairs in Part II. The subjects had a mean consistency score of 14.26 for the first 25 pairs and 23.68 for the last 25 pairs. We observe that the performance of the subjects has increased for the last 25 pairs. This indicates that the subjects were able to assign a talking pattern to each speaker. The consistency scores remained low for the first 25 pairs because the subjects had not yet gained sufficient information on each speaker. For Part III, the difference in the consistency scores for the first and the last 45 pairs was less as prosodic clues were minimized. For the 100 Hz case, the subjects had a mean consistency score of 6.75 for the first 45 pairs, and 9.25 for the last 45 pairs. For the 180 Hz case, the mean consistency score for the last 45 pairs (10.50) was even less than the mean consistency score for the first 45 pairs (11.37).

4.2. Acoustic Feature Transplantations

In this section, we have used PSOLA based methods for transplanting different target speaker characteristics onto the source speech signal to determine the relevance of these characteristics for perception of speaker identity and voice conversion. Four different acoustic features are investigated for their importance in perception of speaker identity:

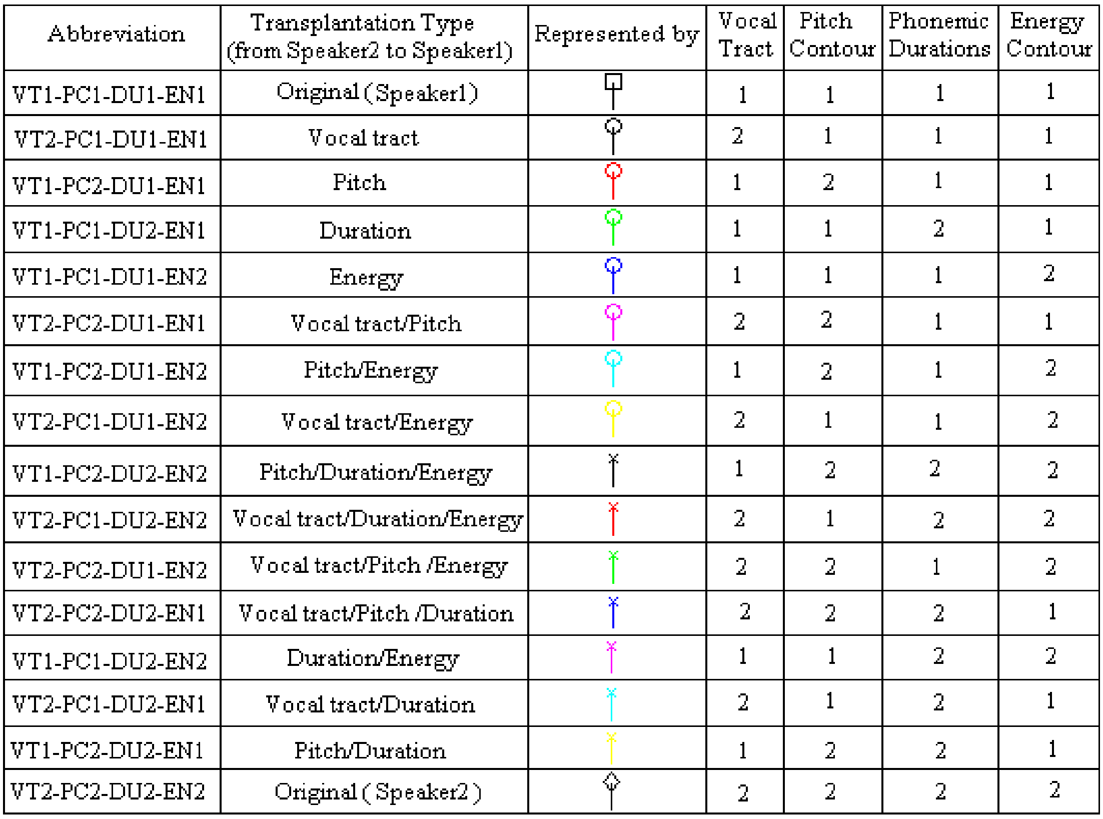
- Vocal Tract (VT)
- Pitch Contour (PC)
- Phonemic Durations (DU)
- Energy Contour (EN)

4.2.1. General Framework

In what follows, we represent these features by the shorthand notations in the parantheses above. We append a number to the right of each feature to show that it comes from either the source speaker (Speaker1) or the target speaker (Speaker2). In our convention, the features follow the order “vocal tract-pitch-duration-energy” and each feature is separated from the others using the “-“ sign. As an example, if we have the vocal tract, pitch and duration information from Speaker1 and the energy contour from Speaker2 (i.e. an energy transplantation), the output is represented as VT1-PC1-DU1-EN2. The procedure for transplantations is as follows:

- Recordings of the same utterances from the source and the target speakers are obtained along with the transcription of the utterance in a separate text file.
- The utterances are automatically labeled and the labels are manually corrected.
- Acoustical feature(s) of Speaker1 is (are) modified to match the acoustic feature(s) of Speaker2 using the phonetic alignment information in a PSOLA based framework.

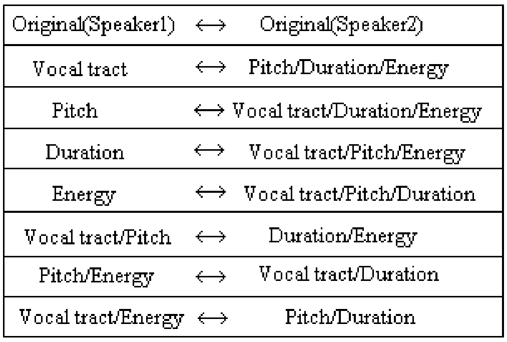
Alignment is performed automatically with a phonetic alignment tool which determines the phoneme boundaries given the Turkish transcription. The labels are manually corrected. Most of the labeling errors occured at plosive sounds like /t/ and /p/. There were also errors in the case of liaison and when the phonemes containing /n/, /m/ or /y/ had extended duration. Note that in liaison two consecutive words are uttered as a single word without having any silence period in between.



Abbreviation	Transplantation Type (from Speaker2 to Speaker1)	Represented by	Vocal Tract	Pitch Contour	Phonemic Durations	Energy Contour
VT1-PC1-DU1-EN1	Original (Speaker1)	□	1	1	1	1
VT2-PC1-DU1-EN1	Vocal tract	○	2	1	1	1
VT1-PC2-DU1-EN1	Pitch	○	1	2	1	1
VT1-PC1-DU2-EN1	Duration	○	1	1	2	1
VT1-PC1-DU1-EN2	Energy	○	1	1	1	2
VT2-PC2-DU1-EN1	Vocal tract/Pitch	○	2	2	1	1
VT1-PC2-DU1-EN2	Pitch/Energy	○	1	2	1	2
VT2-PC1-DU1-EN2	Vocal tract/Energy	○	2	1	1	2
VT1-PC2-DU2-EN2	Pitch/Duration/Energy	×	1	2	2	2
VT2-PC1-DU2-EN2	Vocal tract/Duration/Energy	×	2	1	2	2
VT2-PC2-DU1-EN2	Vocal tract/Pitch /Energy	×	2	2	1	2
VT2-PC2-DU2-EN1	Vocal tract/Pitch /Duration	×	2	2	2	1
VT1-PC1-DU2-EN2	Duration/Energy	×	1	1	2	2
VT2-PC1-DU2-EN1	Vocal tract/Duration	×	2	1	2	1
VT1-PC2-DU2-EN1	Pitch/Duration	×	1	2	2	1
VT2-PC2-DU2-EN2	Original (Speaker2)	◇	2	2	2	2

Table 4.6. All possible combinations of feature transplantations from Speaker2 to Speaker1. “1” denotes that the corresponding feature comes from Speaker1 and “2” denotes that it comes from Speaker2

If we consider the possible number of feature transplantations between Speaker1 and Speaker2, a direct calculation yields 2^4 = 16. Two of these combinations correspond to original utterances of Speaker1 and Speaker2 (i.e. VT1-PT1-DU1-EN1 and VT2-PT2-DU2-EN2). So we are left with 16-2=14 possibilities. These 14 acoustic feature transplantations can be obtained by performing 7 different types of transplantations as described in rows 2-8 of Table 4.6. Each transplantation is repeated by reversing the order of speakers to obtain the rest of the combinations as shown in rows 9-15 in Table 4.6. The first row corresponds to the original utterance of Speaker1 and the last row to the original utterance of Speaker2 in this table. We denote the pair of transplantations that can be obtained by reversing the order of speakers as dual transplantations. Table 4.7 lists the dual acoustic feature transplantations.



Original(Speaker1)	↔	Original(Speaker2)
Vocal tract	↔	Pitch/Duration/Energy
Pitch	↔	Vocal tract/Duration/Energy
Duration	↔	Vocal tract/Pitch/Energy
Energy	↔	Vocal tract/Pitch/Duration
Vocal tract/Pitch	↔	Duration/Energy
Pitch/Energy	↔	Vocal tract/Duration
Vocal tract/Energy	↔	Pitch/Duration

Table 4.7. Dual acoustic feature transplantations

The corresponding time instant in the utterance of Speaker2 is determined by using the information from the labels and the analysis time instant in Speaker1 in Equation 4.2 during PSOLA. Note that we assume a linear time-warping scheme within the phoneme.

$$t_2^i = \frac{(t_1^i - t_1^s)(t_2^e - t_2^s)}{t_1^e - t_1^s} + t_2^s \qquad (4.2)$$

The parameters in Equation 4.2 are as follows:

- i: index of the current label
- t_1^i: time instant in Speaker1 (which is in the i^{th} label) (s.)
- t_2^i: corresponding time instant in Speaker2 (s.)
- t_1^s: start time of the i^{th} label in Speaker1 (s.)

- t_1^e: end time of the ith label in Speaker1 (s.)
- t_2^s: start time of the ith label in Speaker2 (s.)
- t_2^e: end time of the ith label in Speaker2 (s.)

As an example, consider Figure 4.3. This snapshot is taken from VOX where two waveform files are displayed. The waveform on the top is from Speaker1 and at the bottom is from Speaker2. The blue lines show the starting and ending instants of the labels. Current time instant in Speaker1 is shown with a red cursor line as $t_1^i = 0.692$ sec. We want to find the corresponding time instant in Speaker2, t_2^i, using the phonetic labels and Equation 4.2. Note that the starting and ending instants of current labels are also shown in Figure 4.3 as $t^s{}_1$, $t^e{}_1$, $t^s{}_2$, and $t^e{}_2$. For the values given, t_2^i is calculated as 0.435 s. Note that t_2^i is marked with a red cursor line on the waveform for Speaker2. The current feature value for Speaker1 is extracted using t_1^i and appropriate modification is performed on this value to match the corresponding feature value for Speaker2 at t_2^i.

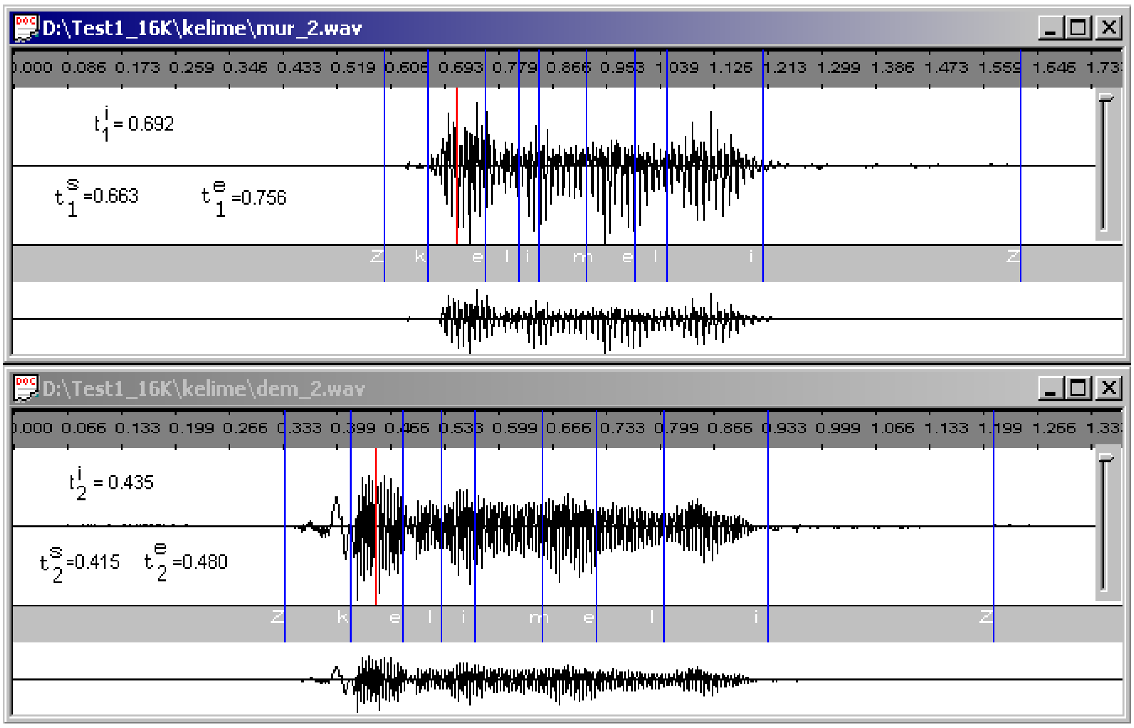


Figure 4.3. Finding the corresponding analysis time instant in Speaker2

4.2.2.Vocal Tract Transplantation

The target vocal tract is transplanted onto the source signal using the corresponding time instant in the target signal. If the target phoneme has extended duration as compared to the source phoneme, the vocal tract vectors should be repeated during synthesis. On the other hand, if the duration of the target phoneme is less than the duration of the source phoneme, appropriate amount of target frames are skipped in PSOLA. We use LSFs as the vocal tract parameters. Once the corresponding target vocal tract parameters are found, the output is synthesized as in Equation 4.3 where w(n) is the coefficients of a Hamming window of size N and x(n) is the current frame obtained from the speech signal pitch-synchronously. Next, we calculate the spectrum H(w) using FFT (Equation 4.4). In Equation 4.4, ceil(k) is a function that rounds k towards +∞. The size of FFT (fftsize) is taken as the minimum power of 2 which is greater than or equal to N. The excitation spectrum E(w) is given by Equation 4.5 where Ps is the vocal tract spectrum estimated using the LSFs.

$$s(n) = \sum_{n=0}^{N-1} x(n)w(n) \tag{4.3}$$

$$H(w) = FFT\{s(n)\}, \ \ fftsize = 2^{ceil(\log_2(N))} \tag{4.4}$$

$$E(w) = \frac{H(w)}{Ps(w)} \ for \ w = 0, \ldots, \ fftsize - 1 \tag{4.5}$$

$$Y(w) = E(w)Pt(w) \ \ for \ w = 0, \ldots, fftsize - 1 \tag{4.6}$$

The output frame FFT, Y(w), is calculated as in Equation 4.6 where Pt is the vocal tract spectrum of the corresponding target frame calculated using the target LSFs. Overlap-add synthesis is then performed to obtain the time domain output. As the vocal tract parameters estimated are the LSFs, we apply preemphasis on the source signal (using a filter $1\text{-}\alpha z^{-1}$) and remove the effect of preemphasis at the output using the inverse preemphasis filter $1/(1\text{-}\alpha z^{-1})$ with α=0.97.

The output of the vocal tract transplantation for a voiced frame is shown in Figure 4.4. The LP spectrum of the transplantation output does not exactly match the spectrum of Speaker2 because the spectrum of the transplantation output is estimated using the synthesized signal.

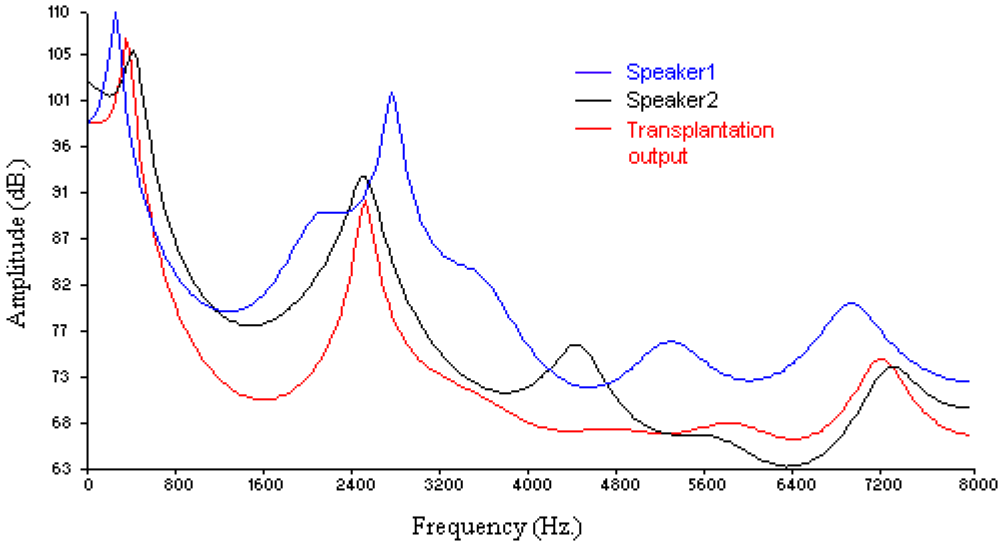


Figure 4.4. Vocal tract spectra for Speaker1, Speaker2, and transplantation output for a voiced phoneme

4.2.3. Pitch Contour Transplantation

We determine the amount of pitch scaling for transplanting the target pitch contour on the signal by using the phonemic time alignment as in the case of vocal tract transplantation. Current pitch value of the source speaker is used with the corresponding target pitch value at the corresponding time index to get the pitch scaling ratio as follows:

$$\beta(t) = \frac{f0_{target}}{f0_{source}} \tag{4.7}$$

where $f0_{target}$ is the instantenous target pitch value, $f0_{source}$ is the instantenous source pitch value and β(t) is the instantenous pitch scale modification factor. We smooth the source and target pitch contours to prevent sudden jumps. We also limit the value of β(t) in the range [0.3, 3.0] in order to avoid exceptionally small or large pitch scaling factors.

In some cases, voiced segments of the source pitch contour will correspond to unvoiced segments in the target contour. We interpolate the target pitch contour linearly in the unvoiced regions. This approach produces satisfactory results as the interpolation is carried out using the pitch values of the target contour at voiced parts. The interpolation of unvoiced regions of the f0 contour is demonstrated in Figure 4.5.

Note that if there is an unvoiced region at the beginning of the utterance, the interpolation is performed by assigning the mean pitch value for the first frame and linearly interpolating the unvocied values until the first voiced frame. If we have an unvoiced region at the end of the pitch contour, we simply replicate the last voiced value. Both situations are demonstrated in Figure 4.5.

Figure 4.6 demonstrates pitch contour transplantation. Note that the output time axis corresponds to the time axis of Speaker1 and the pitch contour of Speaker2 is stretched and shifted on this axis to generate the pitch contour of the transplantation output. Speaker1 is a male and Speaker2 is a female speaker.

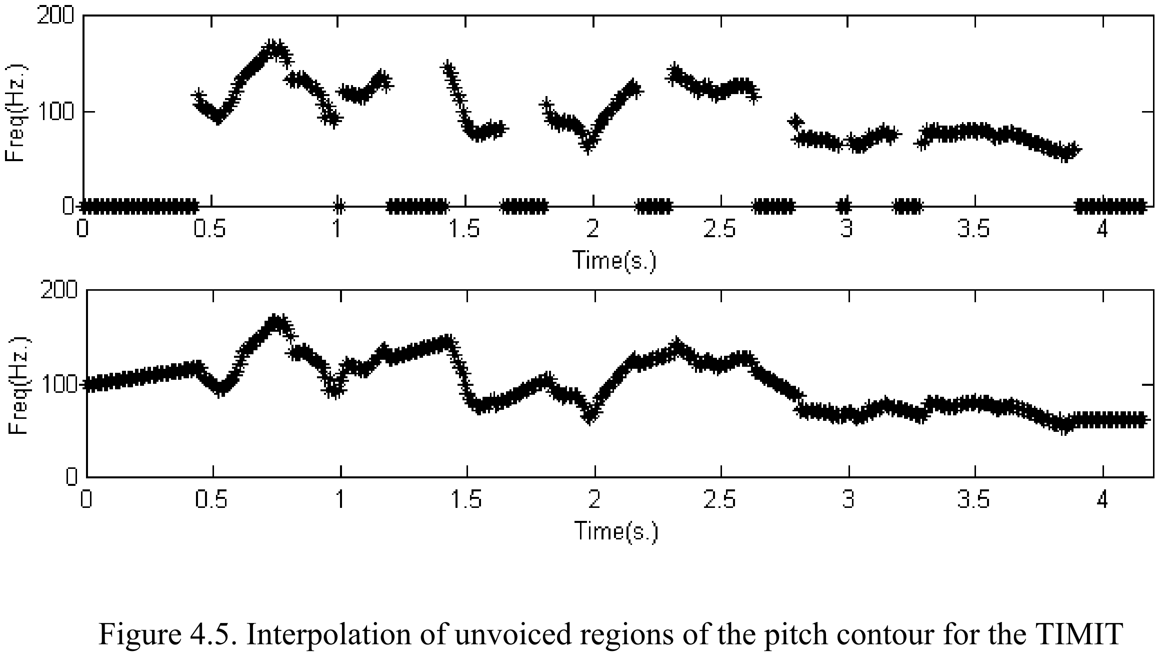


Figure 4.5. Interpolation of unvoiced regions of the pitch contour for the TIMIT sentence of a male speaker “She had your dark suit and greasy wash water all year.”

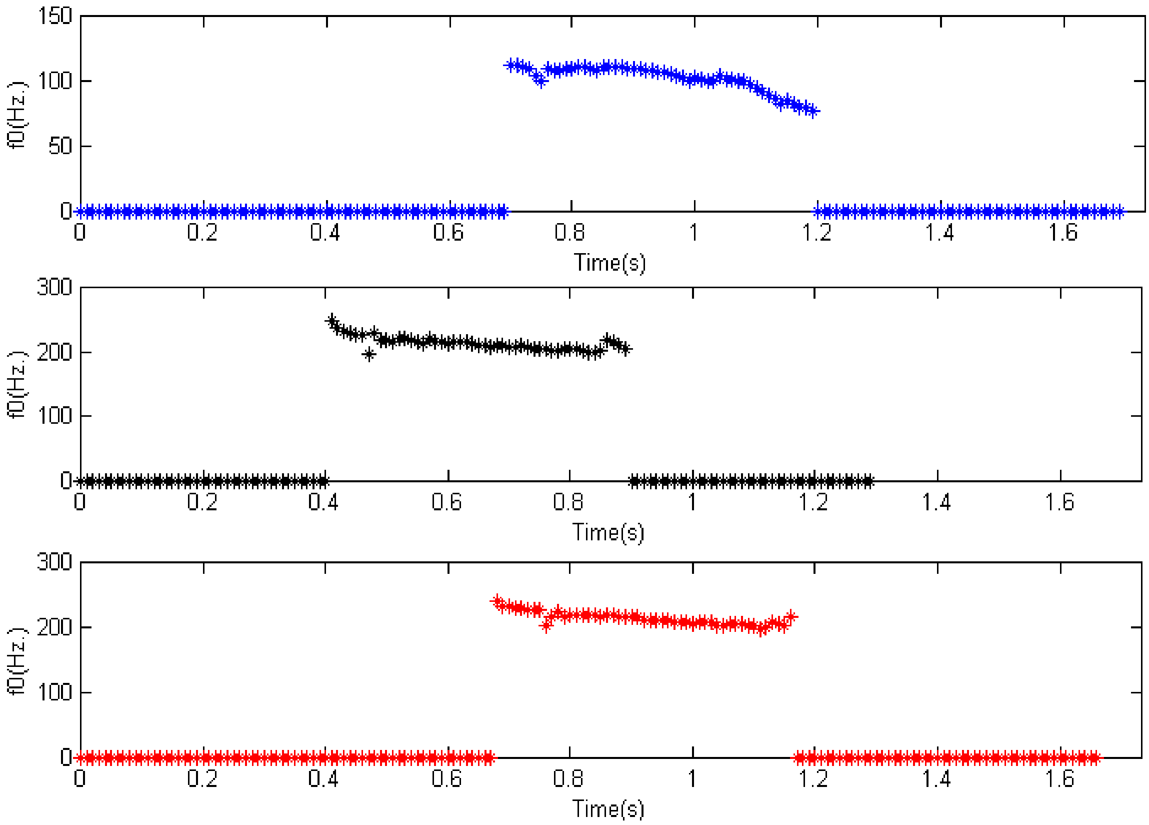


Figure 4.6. Speaker1 (blue), Speaker2 (black), and transplantation output (red)

4.2.4. Transplantation of Phonemic Durations

In transplanting the phonemic durations, we use the original duration information by extracting the excitation signal from the target speaker and modify the rest of the acoustic features to match the source speaker characteristics. This method was preferred because employing TD-PSOLA or FD-PSOLA for phonemic duration modification results in considerable amount of distortion at the output. Both methods are capable of high quality output when the duration scaling ratio does not vary much across neighbouring speech frames. However, there may be drastic changes in the duration scaling ratio of two neighbouring phonemes in the case of phonemic duration transplantation. As the change in the instantenous duration scaling factor increases across phonemes, we get significant distortion at the output. In most of the cases, the output can not be considered as natural. However, we need to have different duration scaling ratios for mimicking the target speaker's duration characteristics in an exact manner. The method we have used does not modify the original target durations but transplants the rest of the features onto the target signal. This is best explained by an example:

Let’s have two speakers, Speaker1 and Speaker2. Let us try to obtain a signal which has the duration characteristics of Speaker2 and the vocal tract, pitch and energy characteristics of Speaker1. The straighforward method will be to estimate the phonemic durations of Speaker1 and Speaker2, and apply appropriate amount of duration scaling across each phoneme of Speaker1 to get the output. However, this results in severe distortion as explained above. The duration scales across phonemes can be smoothed but in this case the output will no longer have the exact duration characteristics of Speaker2. In order to minimize the processing distortion, we follow the alternative path, and modify the vocal tract, pitch and energy contour of Speaker2 to match Speaker1 while not modifying the durations in Speaker2. The output quality is high with the expense of increased computation. This method performs well because the modifications applied to the vocal tract, pitch and energy features do not distort the signal as much as applying duration scaling with a time-varying duration scaling ratio.

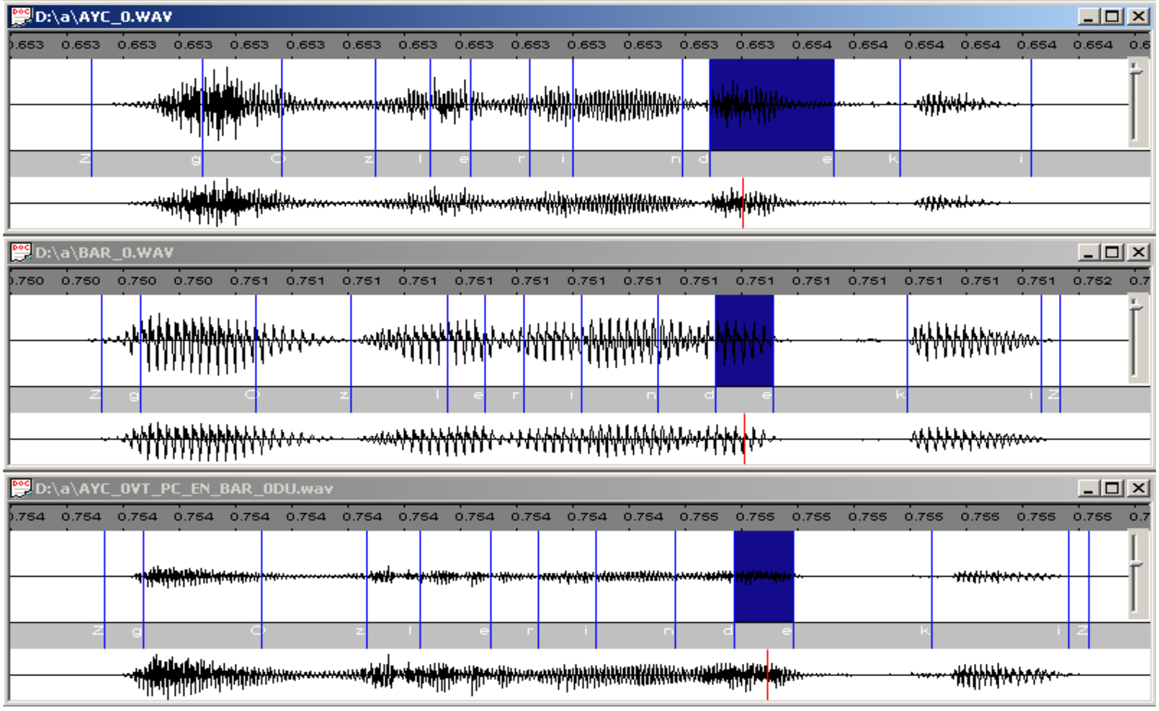

D:\a\AYC_0.WAV
z g O z l e r i n d e k i
D:\a\BAR_0.WAV
z g O z l e r i n d e k i z
D:\a\AYC_0VT_PC_EN_BAR_0DU.wav
z g O z l e r i n d e k i z

Figure 4.7. Speaker1 (top), Speaker2 (middle), output of duration transplantation from Speaker2 onto Speaker1 (bottom)

This method does not lead to any loss of flexibility because we can change the order of speakers to transplant the duration characteristics of Speaker1 on Speaker2. Figure 4.7 demonstrates the results of a phonemic duration transplantation. As an example consider the phoneme /e/ that is shown highlighted in the waveforms. It had rather extended duration in Speaker1 as compared to Speaker2. The output matches the duration in Speaker2.

4.2.5. Energy Contour Transplantation

The energy contour transplantation method is similar to the method used in pitch contour transplantation. Instantenous values of the energy contours are used to obtain the instantenous energy scaling ratio ε(t) as given by Equation 4.8. The energy contours are smoothed before transplantation to reduce discontinuities at the phoneme boundaries. An example is shown in Figure 4.8.

$$\epsilon(t) = \frac{\epsilon_{target}}{\epsilon_{source}} \tag{4.8}$$

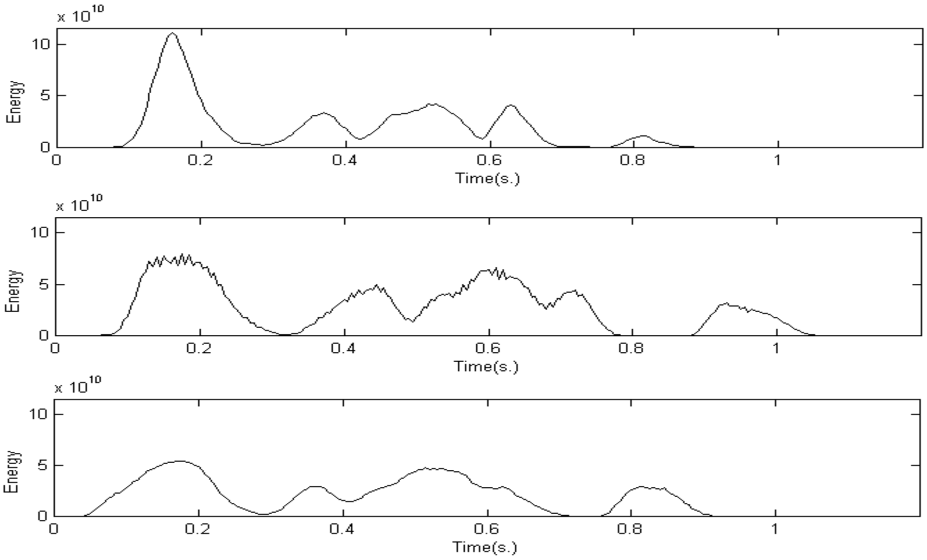


Figure 4.8. Energy contours for Speaker1 (top), Speaker2 (middle), and transplantation output (bottom)

4.2.6. Multi-Feature Transplantations

The rest of the transplantations (Vocal tract/Pitch, Pitch/Energy, and Vocal tract/Energy) are obtained by applying a combination of the four basic methods to transplant the features. As an example, we modify the pitch and the energy contour simultaneously for pitch and energy contour transplantation. When we reverse the order of the speakers, we get all the possible combinations of acoustic feature transplantations.

4.2.7. Methodology

We have used four male and four female speakers for designing the test database. Four types of transplantations are performed as the gender of the speakers are concerned: male-to-male, male-to-female, female-to-female,and female-to-male. The testing procedure is as follows:

- 16 sentences and 16 words are selected randomly from the database for each speaker pair.
- Two sentences and two words are reserved as original signals and will be used for assessing the reliability of the subjects in the tests.
- Each remaining sentence and word is used in one type of transplantation, so we obtain all possible transplantations using a sentence and a word. The subjects were provided with 128 utterance triples. Each triple contained two original utterances from two different speakers and one transplantation output. For reliability measurements, we have used original utterances of either the first or the second speaker as the third item as explained above.
- 10 subjects responded to the triples they have listened by providing one choice and one score. The choice reflected their opinion on the identity of the speaker in the third item. They simply decided whether the third item was uttered by the Speaker1, Speaker2 or none of them. The subjects were also asked to provide a score reflecting how confident they are on their choice of speaker identity. The score was in the range 1-5 on an increasing confidence scale. A score of “1” shows that the subject’s confidence about his/her decision is low meaning that even if the subject has given a decision on the speaker identity, he/she thinks that the third item does

not sound like the identity decided. If the identity choice was “none” then the score reflects the confidence of the speaker about this decision also. The upper limit for the score is “5” which indicates that the subject is confident with his/her choice. Special care is taken for the case of choice “none”. In the case that the transplantation output is similar to both speakers, the subjects were told to assign “none” as the speaker identity with a low confidence score. If the output sounds like a third speaker, the confidence score should be high.

We have used different sentences and words for each case in order to minimize the effect of the acoustical content of the signals and to maximize the capability of listeners to recognize speakers. We observed that when the same recording is used over and over again, the ability of listeners to recognize the speaker identity degrades considerably.

A graphical user interface was designed in VOX to carry out the tests. The subjects were allowed to listen to the sound files as much as they desired. This was the case because we want to evaluate the performance of the subjects on perception of speaker identity when they had sufficient information on the identity of the speakers. We have also used voices of speakers that the subjects were familiar with.

4.2.8. Results

The responses of the subjects are mapped onto a numerical scale to estimate the statistics. We have two scores: the identity score and the confidence score. The identity score is a numerical counterpart of subject decisions on speaker identities and it is obtained by mapping the identity decisions obtained to either “1”,”2”, or “3”. “1” indicates that the transplantation output sounds like Speaker1. “2” shows that the subject can not decide on the speaker identity because it sounds like both speakers or like a third speaker. The subjects were told to assign the lowest confidence score if the output sounds like a third speaker. An identity score of “3” indicates that the perceived identity is Speaker2.

As we reverse the order of the Speaker1 and Speaker2 randomly, the decisions are preprocessed to reflect the similarity to Speaker2. After this preprocessing step, we normalize the scores to cover the range [0.0, 1.0] and estimate the mean and the interquartile ranges for different cases as the gender of the speaker pairs are concerned. The shorthand notations for all the cases are: Overall, M→M, M→F, F→M, F→F, M→?, F→?, ?→M, and ?→F. "Overall" indicates that the statistics are calculated over all gender combinations. M denotes a male speaker, and F a female speaker. The "?" mark indicates that both genders are considered. As an example, F→? denotes the case in which Speaker1 is a female and Speaker2 is either a male or a female. In this case, acoustic features of either a male or a female speaker were transplanted onto a female speaker's utterance.

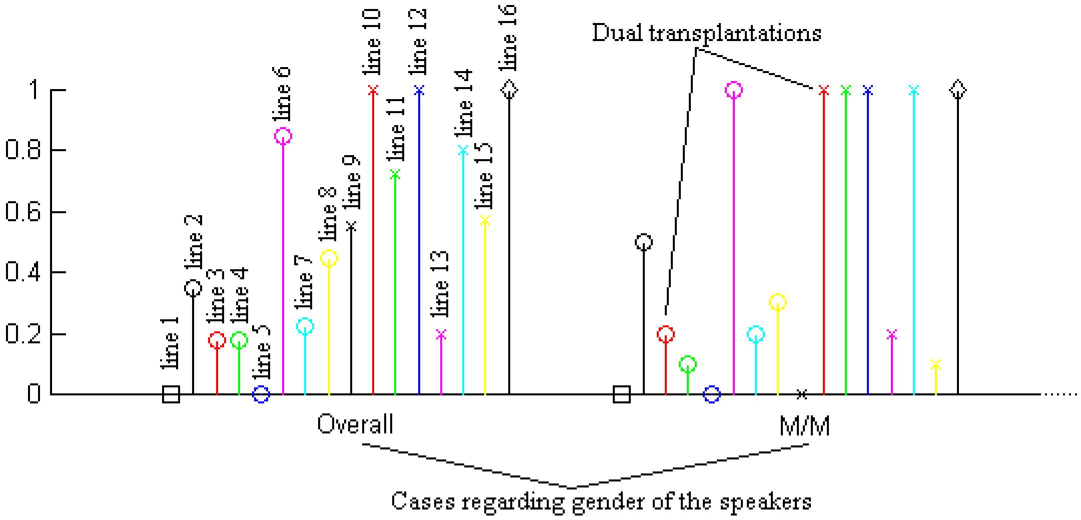


Figure 4.9. Plot for a sample test result

In the figures below, we present the average scores for different cases. In Figure 4.10 and 4.11, the statistics were estimated for all utterances while in Figure 4.12 and 4.13, we estimate them only for words. Figure 4.14 and 4.15 show the mean statistics for sentences. In each figure, we have two subplots. In the first subplot, the average values estimated for different conditions for the identity score are shown. In the second subplot, confidence scores are presented.

In each subplot, we have different group of lines each corresponding to the cases described above regarding the genders. These groups are labeled on the x-axis by the corresponding case. Note that the case "Overall" is included in all plots in order to compare the results of a specific case with the overall trends in which the gender of the speaker pairs are not considered. For each case, we have a group of 16 lines either differing in color or differing on the small mark on their top.

Figure 4.9 shows two group of lines in a sample plot. The line with a small square on the top (line 1) denotes the case when file 3 was exactly the same as file 1 (the source speaker). Theoretically, the identity score should be 0.0 and confidence score should be 1.0 for this case. The results match the theoretical values exactly indicating that the subjects successfully recognized the source speakers. The lines 2-15 correspond to 14 types of transplantations as described above. The last line (line 16) corresponds to the case when the subject was presented the original target recording as the third file. The theoretical identity score for this case is 1.0 with a confidence score of 1.0 and the values observed match these theoretical values exactly. So the target speakers were also identified perfectly.

In the second and third column of Table 4.6, we list the type of transplantations with corresponding lines used in the plots of Figure 4.10 to 4.15 . Note that in each group of lines we have 7 different colored lines with a small circle on the top, each corresponding to the first 7 different type of transplantations. The lines with a small 'x' on the top correspond to the dual of these transplantations as described in Section 4.2.1 and Table 4.7.

In Tables 4.8 and 4.9, we present the interquartile range of the scores for all utterances for the identity score and the confidence score. Note that the interquartile range is defined as the difference of the value which is greater than 75 percent of the data and the value which is greater than 25 percent of the data. So, it is an indicator of the spread of data like the standard deviation.

The possible values of the interquartile range are between 0.00 and the difference between the maximum value and the minimum value of the data. This difference was 1.00 in our case. If the interquartile range is close to 0.00, the data is not wide-spread. This is desired in the case of the transplantation test results because we investigate the general tendency of the subjects for a given transplantation type. Interquartile range values close to 1.00 indicate that the scores are wide-spread. This indicates that the decisions of the subjects are not in agreement for a particular type of transplantation.

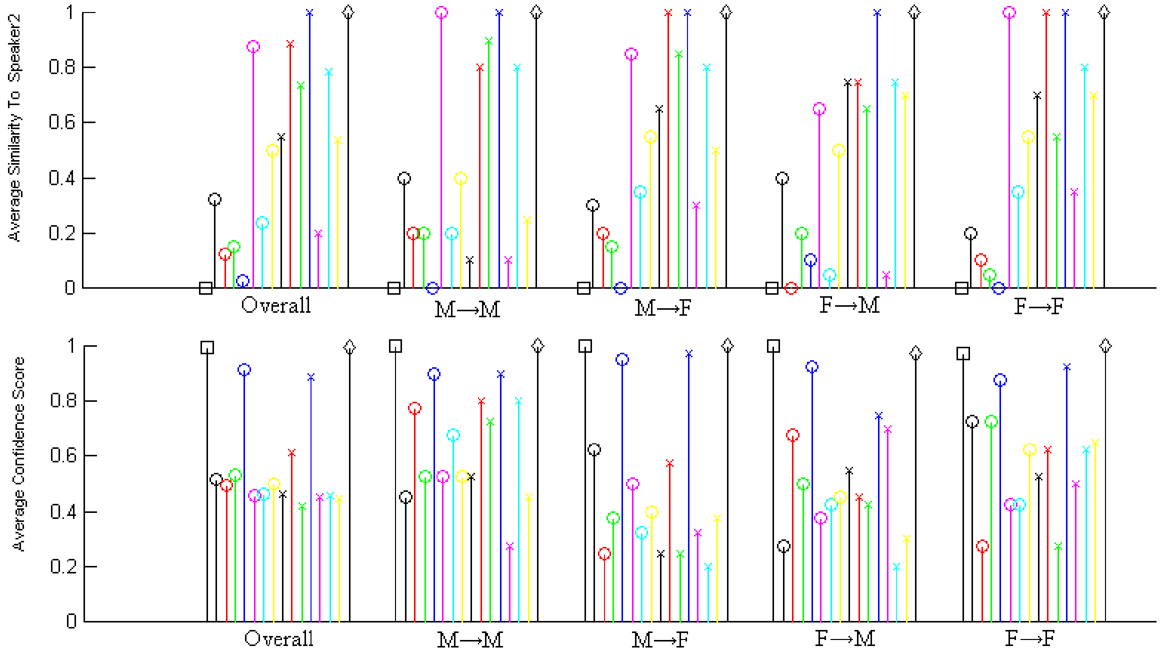


Figure 4.10. Transplantation subjective test results for all utterances

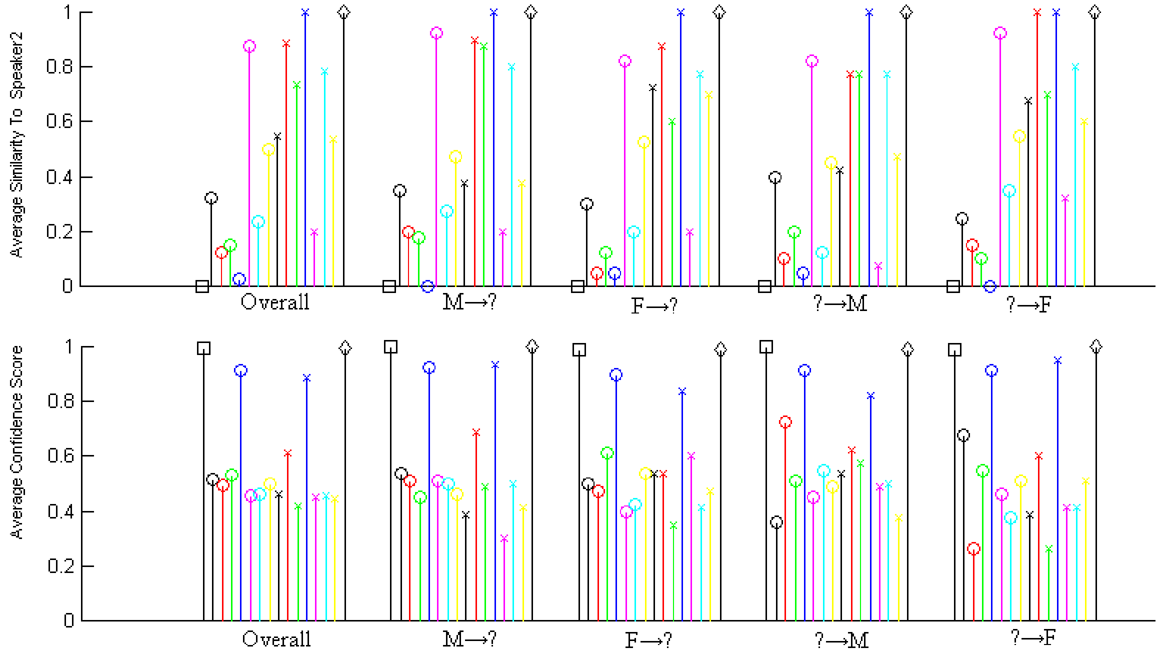


Figure 4.11. Transplantation subjective test results for all utterances

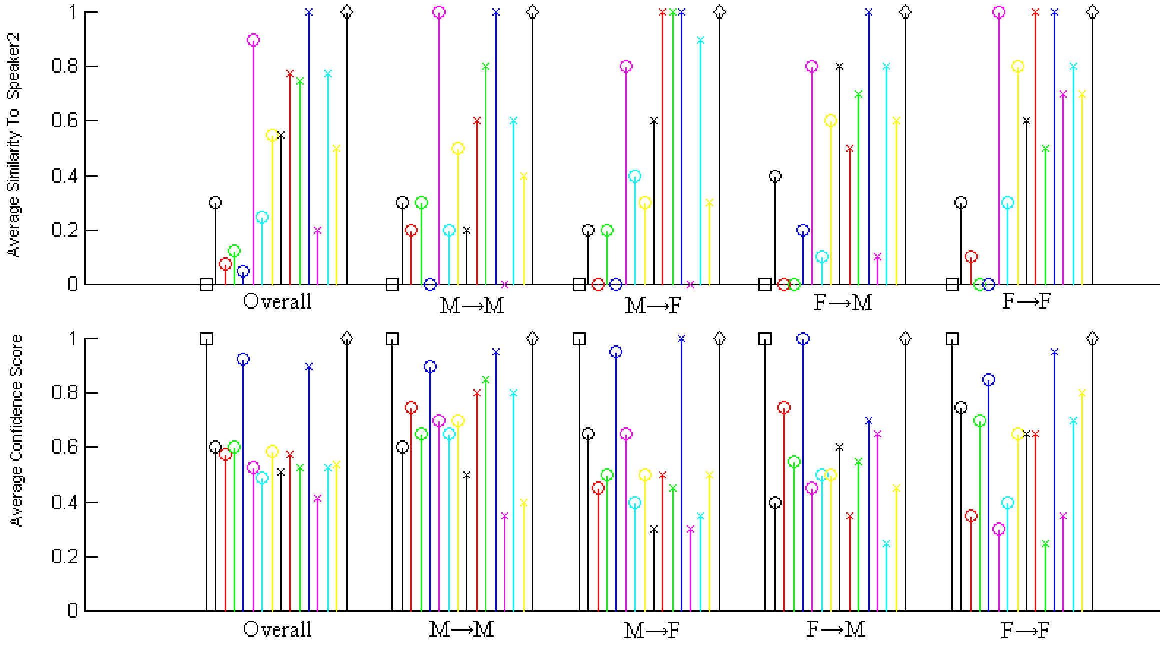


Figure 4.12. Transplantation subjective test results for words

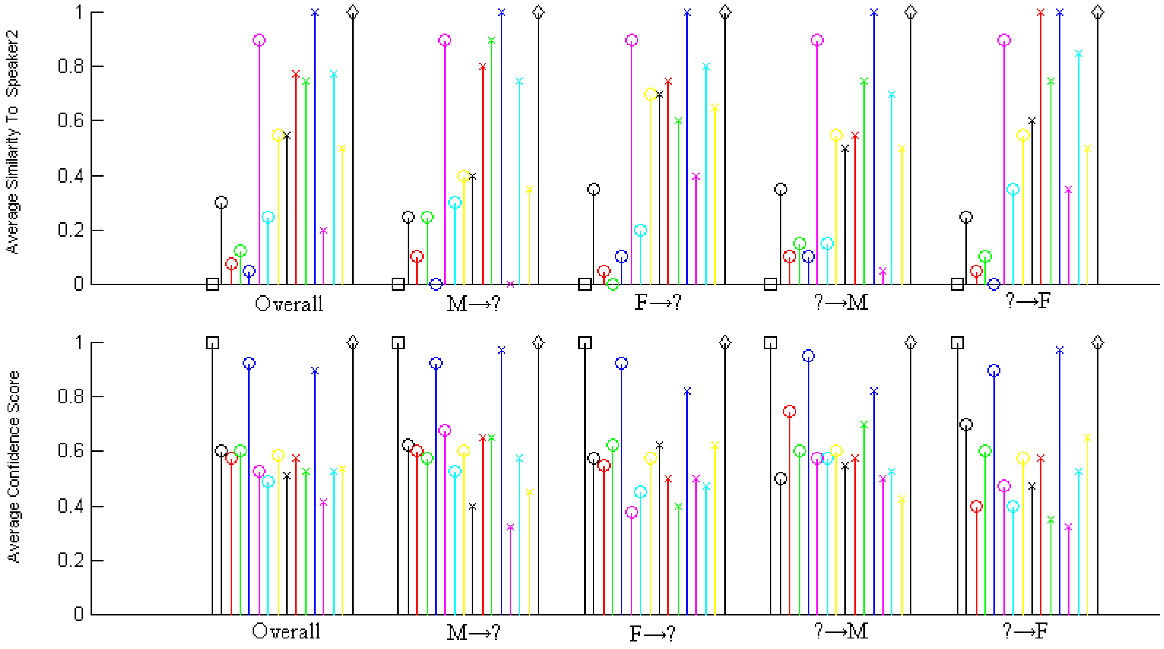


Figure 4.13. Transplantation subjective test results for words

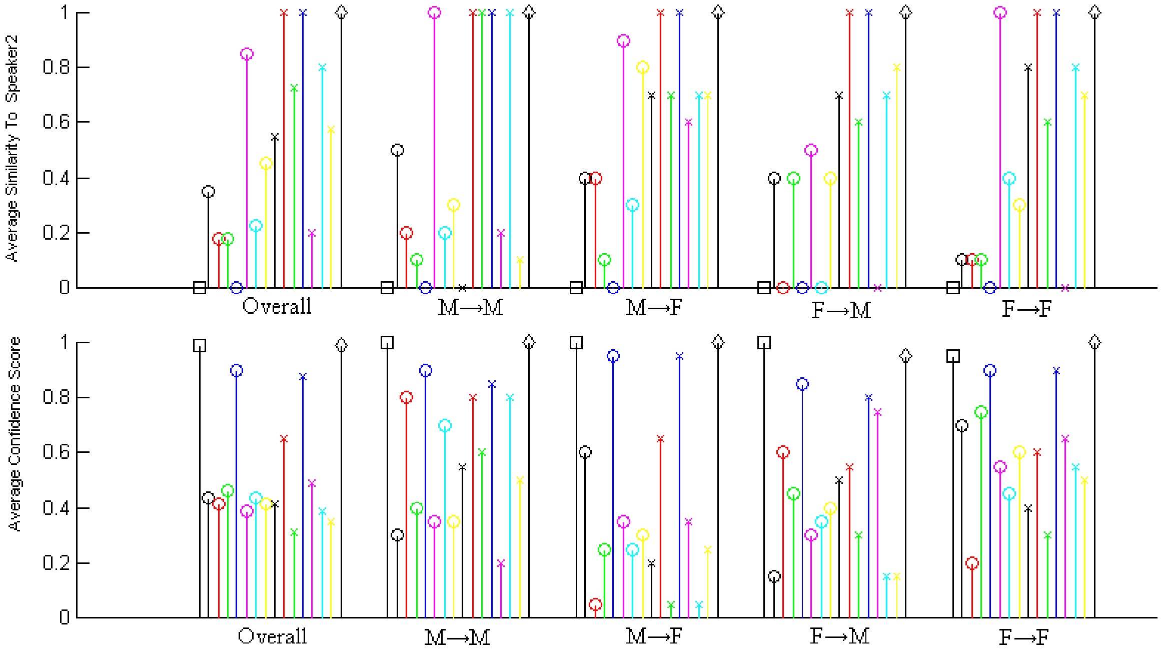


Figure 4.14. Transplantation subjective test results for sentences

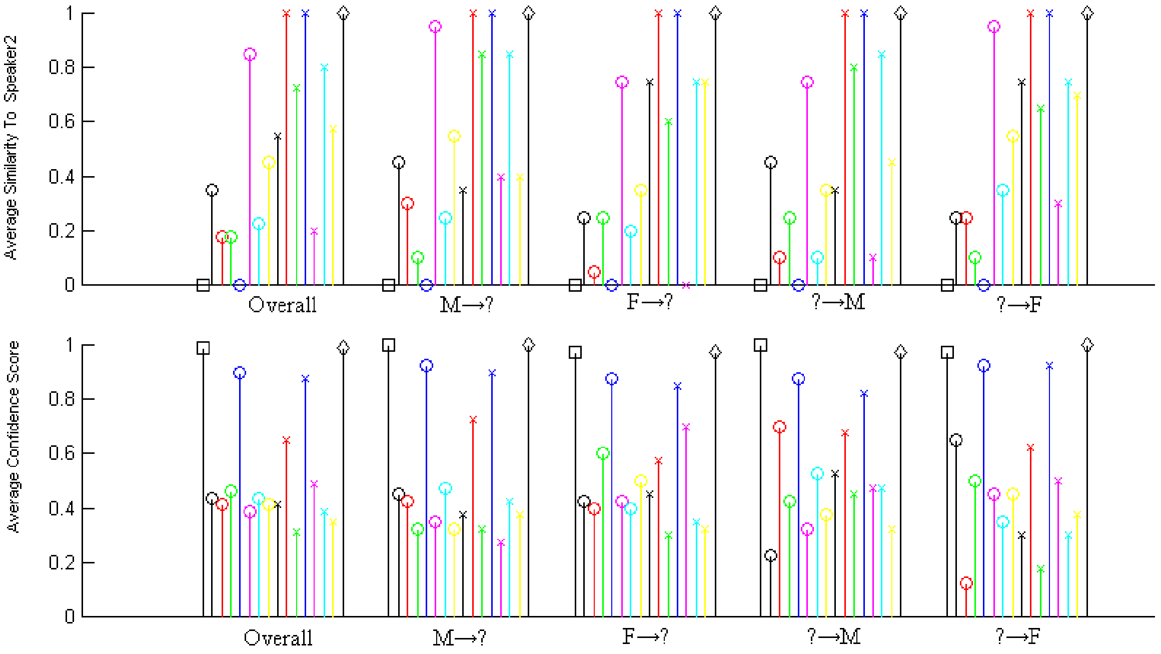


Figure 4.15. Transplantation subjective test results for sentences

Output Type	Overall	M→M	M→F	F→M	F→F	M→?	F→?	?→M	?→F
Original(Speaker1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vocal tract	0.50	1.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Pitch	0.00	0.00	0.50	0.00	0.00	0.50	0.00	0.00	0.50
Duration	0.25	0.50	0.00	0.50	0.00	0.25	0.25	0.50	0.00
Energy	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vocal tract/Pitch	0.00	0.00	0.00	0.50	0.00	0.00	0.50	0.50	0.00
Pitch/Energy	0.50	0.00	0.50	0.00	1.00	0.50	0.25	0.00	0.75
Vocal tract/Energy	1.00	1.00	1.00	0.00	1.00	1.00	0.75	0.50	1.00
Pitch/Duration/Energy	1.00	0.00	0.50	0.50	1.00	0.50	0.50	1.00	0.50
Vocal tract/Duration/Energy	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.25	0.00
Vocal tract/Pitch/Energy	0.50	0.00	0.00	0.50	0.50	0.00	0.50	0.50	0.50
Vocal tract/Pitch/Duration	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Duration/Energy	0.50	0.00	0.50	0.00	0.50	0.50	0.50	0.00	0.50
Vocal tract/Duration	0.25	0.00	0.50	0.50	0.00	0.25	0.25	0.25	0.25
Pitch/Duration	1.00	0.50	1.00	0.50	0.50	0.75	0.50	1.00	0.75
Original(Speaker2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.8. Interquartile ranges of the identity scores for all utterances

Output Type	Overall	M→M	M→F	F→M	F→F	M→?	F→?	?→M	?→F
Original(Speaker1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Vocal tract	0.62	0.50	0.50	0.75	0.50	0.75	0.50	0.75	0.50
Pitch	0.75	0.50	0.75	0.25	0.75	0.88	0.50	0.50	0.75
Duration	0.50	0.75	0.75	0.50	0.25	0.62	0.25	0.62	0.50
Energy	0.25	0.25	0.00	0.00	0.25	0.12	0.25	0.12	0.25
Vocal tract/Pitch	0.38	0.50	0.50	0.25	0.25	0.50	0.25	0.38	0.38
Pitch/Energy	0.50	0.25	0.50	0.50	0.50	0.62	0.50	0.50	0.62
Vocal tract/Energy	0.50	0.25	0.75	0.75	0.25	0.50	0.25	0.38	0.50
Pitch/Duration/Energy	0.50	0.00	1.00	1.00	0.00	0.62	0.62	0.62	0.62
Vocal tract/Duration/Energy	0.75	0.25	0.50	0.75	0.75	0.50	0.50	0.62	0.62
Vocal tract/Pitch/Energy	0.62	0.50	0.25	0.50	0.75	0.50	0.75	0.62	0.75
Vocal tract/Pitch/Duration	0.25	0.25	0.00	0.50	0.25	0.00	0.25	0.25	0.00
Duration/Energy	0.50	0.50	0.25	0.50	0.75	0.50	0.62	0.50	0.62
Vocal tract/Duration	0.50	0.25	0.50	0.25	0.25	0.62	0.38	0.50	0.62
Pitch/Duration	0.75	0.50	0.75	1.00	0.25	0.75	0.88	0.75	0.62
Original(Speaker2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.9. Interquartile ranges of the confidence scores for all utterances

As the overall trends are considered, the most convincing transplantation is obtained when the vocal tract, pitch and durations are transplanted. The average confidence scores are also high for this type of transplantation. The interquartile ranges are low as shown in Tables 4.8 and 4.9, so most of the subjects provided similar (high) scores for this type of transplantation. The dual transplantation –i.e. transplantation of the energy contour- was rated as the least convincing one regarding the similarity to Speaker2 in most of the cases. Pitch contour and Pitch/Energy contour transplantations had lower scores also.

In the case of the words, these scores were closer to the scores obtained for energy contour transplantation indicating that pitch information is not as important as in the case of sentences. This is expected because prosodic characteristics are more variant in sentences. Vocal tract/Pitch contour transplantation, Vocal tract/Duration/Energy transplantation and Vocal tract/Duration transplantation also had high scores.

In the case of single feature transplantations, vocal tract appears as the most relevant feature. In most of the cases, pitch and duration had similar scores. The least relevant feature was the energy contour. As the gender of the speakers is considered, pitch transplantations had relatively higher scores when the genders of the two speakers were different.

As the dual transplantations are considered, the average scores indicate that if any transplantation is assigned a higher score, the dual transplantation gets a lower score. As an example consider the energy contour transplantation and the vocal tract/pitch/duration transplantation in Figure 4.10 for the overall case (i.e. first group of lines in the plot). We observe that the average similarity to the target speaker is lowest for energy contour transplantation (blue line with a 'o' sign on top). The dual case which corresponds to vocal tract/pitch/duration transplantation (blue line with a 'x' sign on top) was rated with a high score. It is clear that if transplanting a subset of the features does not produce an output that sounds like the target speaker, transplanting the rest of the features or including more features will have a better chance.

The importance of these transplantations stems from the following facts:

- Transplantations provide the framework for evaluating the importance of any feature combination in terms of perception of speaker identity.
- Transplantations demonstrate the theoretical limits for conversion of any feature. As an example, consider the case of pitch transplantation. It corresponds to the ideal

case for voice conversion, i.e. the case that we already have the natural pitch contour of the target speaker at hand and carry out pitch scaling accordingly. So any transplantation provides the means to evaluate the voice conversion output in which the feature(s) is (are) automatically modified by the voice conversion algorithm.

5. NEW METHODS FOR VOICE CONVERSION

This chapter introduces three new methods that can be used in the design of voice conversion algorithms. The first two methods described in Sections 5.1 and 5.2 employ subband processing to estimate and modify the vocal tract characteristics considering the perceptual properties of the human auditory system. In Section 5.3, we present the comparison of the new vocal tract transformation methods with the full-band method. The last method is described in Section 5.4 and it can be used for transformation of the pitch contours.

5.1. DWT System for Subband Based Voice Conversion

In this section, we describe the framework for a subband based voice conversion system that makes use of the DWT (Turk and Arslan, 2002). As described in Section 3.4, STASC algorithm operates on the full-band spectrum in both training and transformation stages. However, the subjective tests presented in Section 4.1 demonstrate the dependence of perception of speaker identity on the spectral distribution of the signals. Thus, it will be useful to design a voice conversion method that takes these perceptual characteristics into account. For this purpose, we modify the training procedure described in Section 3.4.1 by estimating the spectral parameters from the subbands. In the transformation stage, different subbands are processed separately in order to reduce distortion and increase computational efficiency. The method presented in this section produces satisfactory results at a sampling rate of 44.1KHz as demonstrated by the subjective tests. The opportunity to perform voice conversion at high sampling rates is required for dubbing and looping applications.

Modeling the source and the target spectra in detail becomes tedious as the sampling rate increases. The number of parameters should be high enough to represent the spectrum accurately. It is common practice to use 16-18 LSFs at 16 KHz. When the sampling rate is higher than 16 KHz, the number of sufficient LSFs increase. However, in some cases, when we increase the order beyond 18, we get significant distortion in transformation quality using the full-band based method due to several reasons:

- Accuracy of the root-finding algorithms used for calculating LSFs degrade for higher order polynomials that must be solved in the case of higher prediction orders.
- Interpolation of the LSFs for the full-band spectrum as defined by Equation 3.33 may lead to formant shifts. This is due the fact that LSFs are pure frequency values and as the range they cover increases, linear interpolation results in more shifts in LSF pairs. This causes the formants to shift.
- Modification of higher frequency bands may degrade output quality. These regions of the spectrum contain non-speech components that cause distortion when modeled and transformed like speech.

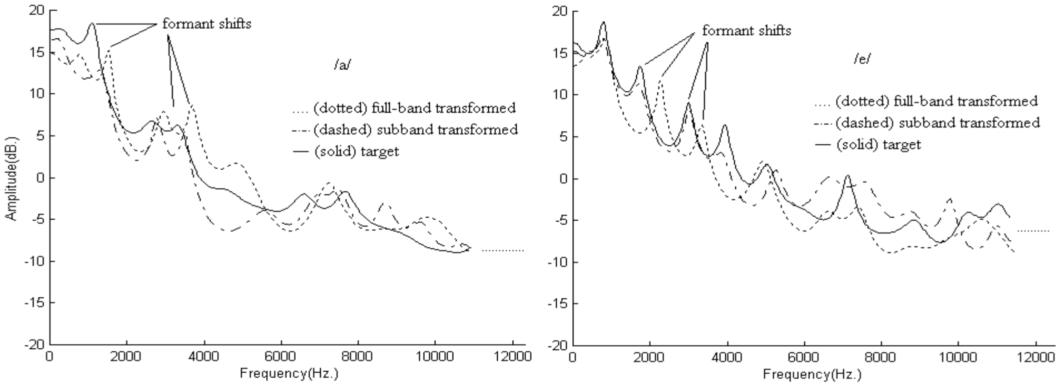


Figure 5.1. Full-band vs. subband based voice conversion at 44.1KHz. (Whole spectrum is not displayed)

As an example, consider Figure 5.1. In the case that we use 18th order LPC to represent the spectrum at 44.1 KHz, the transformed speech had a child-like voice quality. This is due to shifts in formants towards higher frequencies by interpolation as given in Equation 3.33 in Section 3.4.2. When the LSFs are interpolated in the baseband range (i.e. 0-5512.5 Hz), the amount of shift in the formants will be less as compared to the full-band range (i.e. 0-22050 Hz). This is shown in Figure 5.1. In this figure, the source is a male speaker and the target is a female speaker. The speech segments correspond to /a/ and /e/ in the utterance “It’s h/a/rd to t/e/ll an original from a forgery” respectively. The spectrum for the full-band based conversion (dotted plots) possesses severe shifts of formants below 5KHz which cause a child-like voice quality.

In general, training is computationaly the most intensive process in voice conversion especially when the source and target acoustical spaces are well covered using sufficient amount of training data. Using the DWT based method, the computational load is reduced as lower prediction orders can be used. The codebook based transformation procedure relies on codebook search basically. We search the closest parameter vectors to the incoming speech frame parameters in the source codebook. The corresponding target parameters are obtained from the target codebook. As the number of parameters in the codebook increases, the search from the codebook (which contains typically 1000 - 5000 speech units) to find the closest entries takes more time. However, employing the DWT based method, transformation takes less time as sampling rate is reduced. This is due to the fact that at a lower sampling rate the sufficient number of vocal tract parameters are reduced. Following sections describe the DWT based training and transformation stages in more detail.

5.1.1. Training

In the training stage, separate codebooks are generated for each subband to be transformed. We have observed that using only the lower subbands for Sentence HMM based alignment produces satisfactory alignment performance. In our experiments, we have reduced the sampling rate up to 11 KHz and obtained satisfactory results. This is expected since speech signals are restricted to lower frequencies due to the physical properties of the human speech production system. Most of the formants that contribute to perception of speaker identity reside in the frequency band [0, 5.5] KHz. In Section 4.1, we have shown that this frequency range was very important for perception of speaker identity.

The flowchart for the subband based training algorithm is shown in Figure 5.2. First, source and target training utterances are decomposed using DWT. Using the subband signals for automatic alignment, the codebooks are generated which contain the acoustical parameter mapping between the source and target speakers. For 44.1 KHz recordings, we have used four subbands each covering 5.5 KHz frequency range. The first subband covers approximately all the speech components to be used in the generation of the codebooks. The rest of the subbands cover the frequency range 5.5

KHz-22 KHz. We employ only FD-PSOLA based prosody modifications in these subbands. As the sampling rate is reduced to 11025 Hz, training takes much shorter time. It is possible to obtain a detailed spectral representation using even 16th order LP analysis. Subband codebooks for both speakers are generated using the alignment information similar to the full-band case. The codebooks contain LSFs, f0, energy and duration as acoustical information. Although extra processing must be carried out for subband decomposition, the time required for the overall training process is reduced considerably. This is due to the fact that DWT based decomposition can be realized fully in time domain using FIR filters as described in Section 3.7.

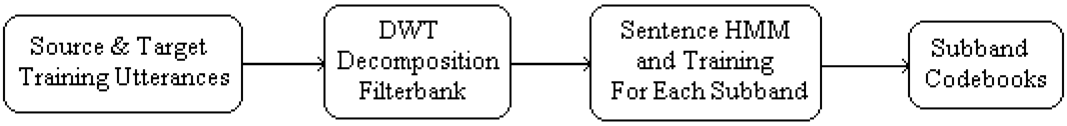

Source & Target Training Utterances
DWT Decomposition Filterbank
Sentence HMM and Training For Each Subband
Subband Codebooks

Figure 5.2. DWT based training algorithm

5.1.2. Transformation

The flowchart for the subband based transformation algorithm is shown in Figure 5.3. The input speech signal from the source speaker is first decomposed using the DWT filterbank. The subband signals are processed separately for transforming the vocal tract characteristics from the source speaker to the target speaker in the subband domain. The vocal tract transformation process is a frequency domain filtering operation in which the source vocal tract spectrum is transformed to the target vocal tract spectrum using the codebook entries as described in Section 3.4.2. Post-filtering may be applied optionally for removing audible noise in different frequency bands. In most of the cases, attenuating high frequencies produce better results. We have preferred another method that does not modify the spectral envelope at high frequency subbands. We have noticed that transforming higher frequencies does not contribute much to the quality of the transformation but increases distortion as non-speech components are modeled and transformed as speech. In this case, we have applied bandpass filtering at the subbands that will not be transformed.

The subbands are separated into two groups for reconstruction. The subbands with modified vocal tract characteristics (i.e. the lower subbands) are input to the DWT reconstruction filterbank. This filterbank is provided with zeros at the non-modified subbands. By this way, we reconstruct the vocal tract modified signal. The rest of the spectrum is obtained by summing the vocal tract modified signal with the bandpass filtered signals. The prosodic modifications are performed at the second pass after the vocal tract is modified. We have used the STASC framework described in Section 3.5 and 3.6 for prosodic modifications. The excitation spectrum is processed in the full-band range.

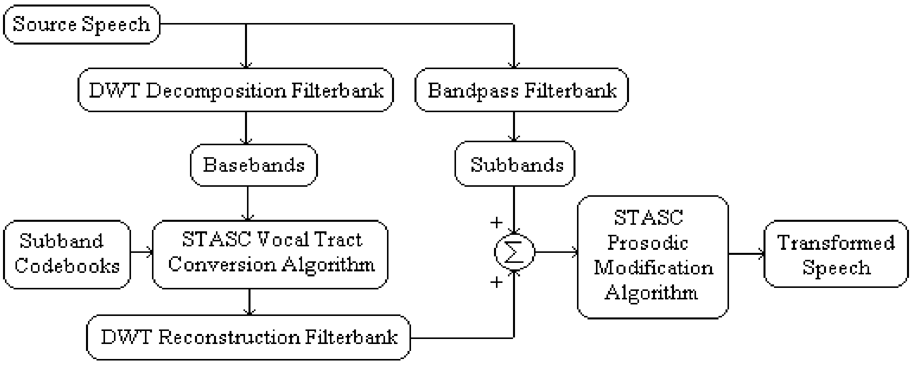

Source Speech
DWT Decomposition Filterbank
Bandpass Filterbank
Basebands
Subbands
Subband Codebooks
STASC Vocal Tract Conversion Algorithm
+
Σ
+
STASC Prosodic Modification Algorithm
Transformed Speech
DWT Reconstruction Filterbank

Figure 5.3. Subband based transformation algorithm

In Section 5.3, we present several spectral plots for vocal tract transformation using DWT. The subjective tests of Section 6.2 compare the performance of the full-band and the DWT based subband voice conversion systems. Section 6.3 presents more detailed subjective tests for the comparison of three vocal tract conversion methods.

5.2. Selective Preemphasis System

In this section, we combine the motivation for preemphasis with perceptual subband processing to estimate the vocal tract spectrum in detail. The new method is evaluated in terms of objective and subjective measures in Chapter 6. This spectral estimation method is used as part of the vocal tract conversion system for increasing the resolution at each subband and converting the vocal tract in more detail. We refer to this new method as selective preemphasis.

Although the DWT based system described Section 5.1 provides efficient solutions at higher sampling rates, it has disadvantages. When the aim is to perform modification in all subbands, aliasing distortion causes a reduction in the output quality. There are two reasons for this kind of distortion:

- The reconstruction filterbank does not provide perfect reconstruction because each subband is a modified version of the original.
- If a spectral peak occurs at one of the subband boundaries, it is not estimated and transformed accurately.

For these reasons, we have investigated a new method that provides the means to model and transform different frequency regions in different amounts of spectral detail with less distortion and more flexibility. As described in Section 3.2.4, the LSFs cluster close to the peaks in the spectrum. This property has led to the preemphasis methods to improve the numerical properties of LPC analysis prior to the estimation of the LSFs. We have observed that more LSFs are devoted to the bandpass region of the spectra by applying LP analysis for bandpass filtered spectra. The bandstop regions of the spectrum contain less number of LSFs as compared to the bandpass regions. This is illustrated for several bandpass filtered signal spectra in Figure 5.4.

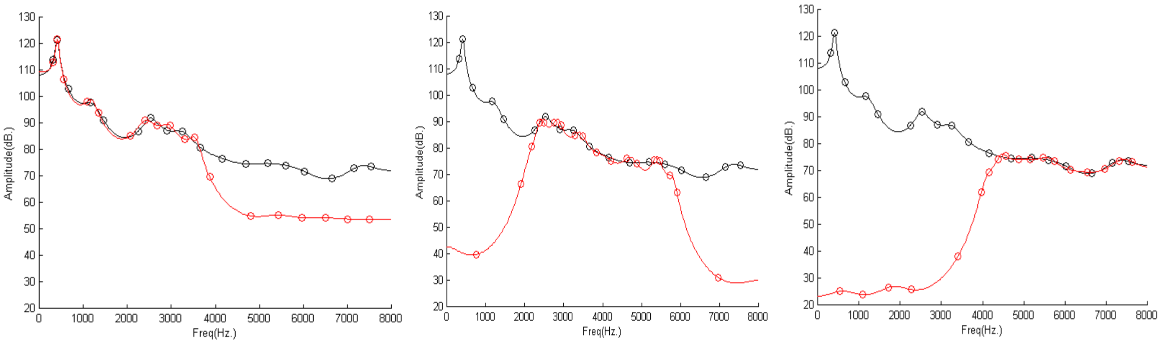


Figure 5.4. Effect of bandpass filtering on the LSF locations

In Figure 5.4, LSFs are marked with small circles on the spectrum. They are calculated using the LP coefficients obtained in the corresponding subband. Black plots are the original spectral estimates using a prediction order of 18 (for 16 KHz). In the first plot (leftmost), 10 LSFs fall in the region 0-4KHz whereas the number of LSFs in the range 0-4 KHz increase to 12 when a bandpass filter having the bandpass region at 0-4 KHz is used. In the second (middle) plot, the signal is filtered with a bandpass filter with a pass-band between 2-6 KHz. Note that the number of LSFs increases from 10 to 16 in the bandpass region in this case. The last figure (rightmost) indicates an increase from 8 to 13 LSFs in representing the region between 4-8 KHz. Note that we have a total of 18 LSFs as the prediction order is 18 for all the cases. It is clear from Figure 5.4 that by bandpass filtering, it is possible to enforce the analysis process to model a specific region of the spectrum in more detail. In this section, we describe the general framework for a new subband based spectral analysis and synthesis algorithm that performs detailed spectral estimation using linear prediction.

5.2.1. Analysis

The flowchart of the analysis algorithm is shown in Figure 5.5. It can be used to estimate the vocal tract and the excitation spectrum as follows:

- The speech signal s(k) is bandpass filtered with N bandpass filters in the perceptual filterbank to obtain the subband signals.
- Each subband signal is processed frame-by-frame. We denote the ith windowed speech frame by $\mathbf{s^i_w}$. (A Hamming window is used)
- LP analysis is performed at each subband to obtain the nth LPC coefficient vector for the ith frame of the analysis stage which is denoted by $\mathbf{a^i_n}$. Note that n is the subband index, and i is the frame index.
- LP spectrum vector $\mathbf{P^i_n}$ is calculated for each subband using the corresponding LP coefficient vector $\mathbf{a^i_n}$.
- The vocal tract spectrum $H^i(w)$ is estimated from the subband LP spectra using the following formula:

$$H^i(w) = \sum_{n=1}^{N} c_n(w) P_n^i(w) \; for \; w = 0, \ldots, W-1 \tag{5.1}$$

where W is the FFT size (which was dependent on the pitch-synchronous analysis frame size in our case), n is the subband index, and i is the frame index.

- The weight of the LP spectrum of a subband at a specific frequency w is denoted by $c_n(w)$ and it is given by:

$$c_n(w) = \begin{cases} \dfrac{w - w_1}{w_1 - w_2} + 1, \ if w_1 \leq w \leq w_2 \\ 0, \ else \end{cases} \tag{5.2}$$

where w_1 : lower cut-off frequency of the $(n+1)^{th}$ bandpass filter
w_2 : higher cut-off frequency of the n^{th} bandpass filter

The condition $w_1 \leq w \leq w_2$ ensures that the bandpass regions of the neighboring filters in the filterbank have a specific amount of overlap in the frequency domain. This overlap ensures better spectral estimation at the bandpass filter boundaries. If bandpass filters without overlap are used, the spectral estimation performance at the subband boundaries will degrade as in the case of DWT. It is possible to use sharp cut-off filters to prevent this situation. In this case, the order of the filters should be high and the computational load is increased. The interpolation procedure described above provides reliable spectral estimates at the subband boundaries as it uses information from both of the subbands in the overlapped region. $c_n(w)$ given by Equation 5.2 is in fact a linear interpolation rule at the overlapping region. It is also possible to use a different interpolation scheme.

- The original speech signal is processed frame-by-frame using windowing. FFT spectrum $\mathbf{S^i_w}$ is estimated using the original signal for the i^{th} frame.
- The excitation spectrum for the i^{th} frame is given as:

$$E^i(w) = \frac{S^i(w)}{H^i(w)} \tag{5.3}$$

The vocal tract spectrum and/or excitation spectrum can be processed separately after the analysis procedure. We describe the application of this method for voice conversion in Sections 5.2.5. and 5.2.6. However, this is a general analysis-by-synthesis scheme which can be used in different applications.

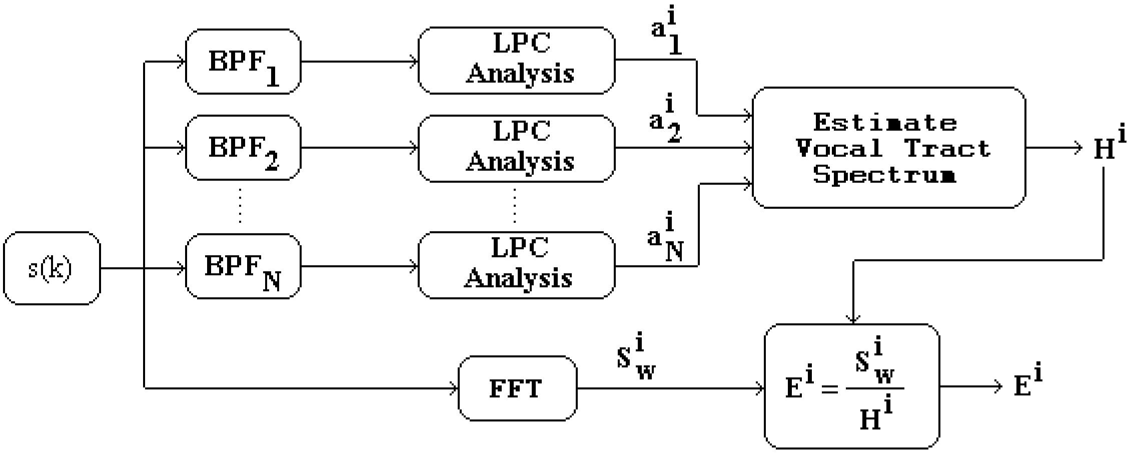


Figure 5.5. Analysis algorithm for spectral estimation using selective preemphasis

5.2.2. Synthesis

In the synthesis stage, we use the synthesis LP coefficient vectors to reconstruct the vocal tract spectrum using Equation 5.1. The synthesis LP coefficients can be a modified version of the analysis coefficients depending on the application. For example, they are the target LP coefficients estimated from codebooks for voice conversion. The synthesis algorithm is shown in the flowchart of Figure 5.6 and it proceeds as follows:

- Use the synthesis LP coefficients for the i^th^ frame and n^th^ subband, $\mathbf{a}^{\mathbf{i}}_{\mathbf{n}}$, to estimate the synthesis vocal tract spectrum vector $\mathbf{H}^{\mathbf{i}}$. The formulas given in the analysis stage are used by replacing the analysis parameters with the synthesis parameters. The synthesis LP coefficients and the synthesis LP spectrum calculated from these coefficients are employed.
- Obtain the synthesis FFT spectrum for the i^th^ frame by multiplying the synthesis vocal tract spectrum by the synthesis excitation spectrum. The synthesis vocal tract spectrum can also be a modified version of the analysis excitation spectrum depending on the application. For voice conversion, the analysis excitation spectrum is modified using FD-PSOLA.

In the following subsection, we demonstrate the selective preemphasis based spectral estimation method using a simple bandpass filterbank. This example demonstrates that detailed spectral analysis at a low LP order is possible with this method.

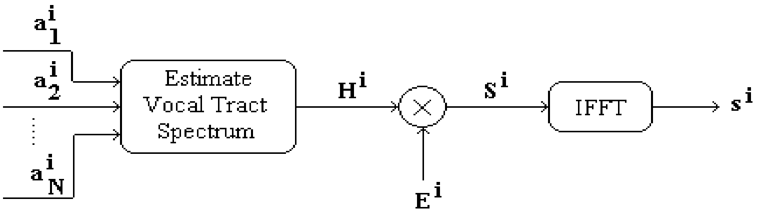


Figure 5.6. Synthesis algorithm for spectral estimation using selective preemphasis

5.2.3. Demonstration

The equally spaced bandpass filters in Figure 5.7 are used for demonstration purposes. For voice conversion, we have employed a perceptual bandpass filterbank and used higher resolution in lower frequencies (i.e. the bandwidths increased with the increasing center frequency logarithmically).

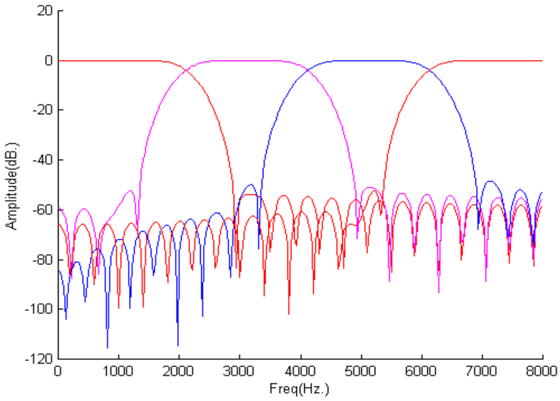


Figure 5.7. Bandpass filterbank with equally spaced bands for demonstration

Figures 5.8, 5.9, and 5.10 present the selective preemphasis based spectral estimation. The TIMIT utterances “She had your dark suit and greasy wash water all year” by a male and a female speaker was used for demonstration purposes. A prediction order of 18 was used both in the LP and selective preemphasis based spectral estimation. The sampling rate was 16 KHz. Each colored spectrum plot corresponds to a subband. We observe that spectral estimation is more detailed and spectral nulls are betted modeled using the selective preemphasis method.

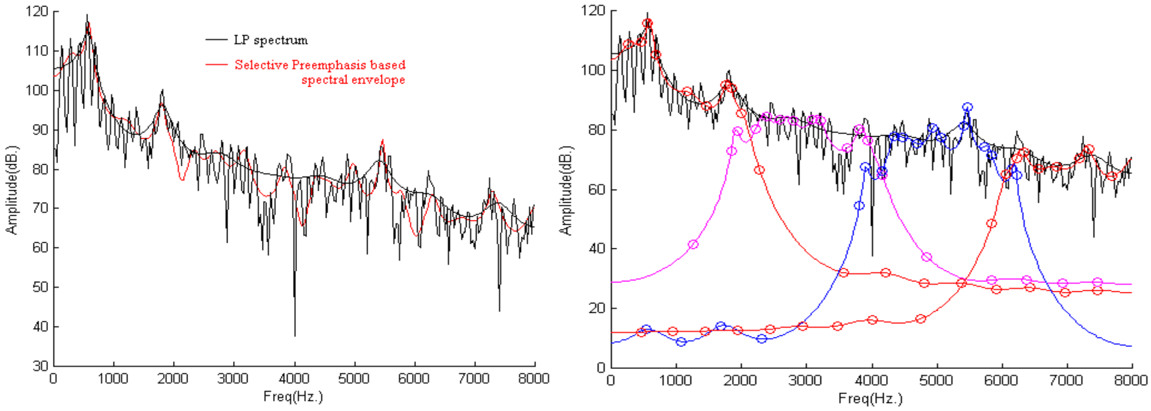


Figure 5.8. LP vs. selective preemphasis based spectral estimation (left). Spectral estimation process from subbands (right) for the phoneme /a/ (male speaker)

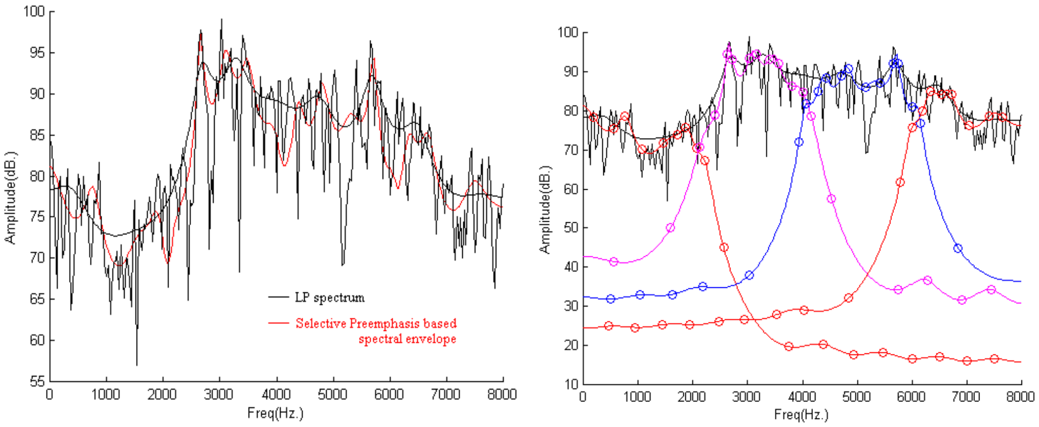


Figure 5.9. LP vs. selective preemphasis based spectral estimation (left). Spectral estimation process from subbands (right) for the phoneme /sh/ (male speaker)

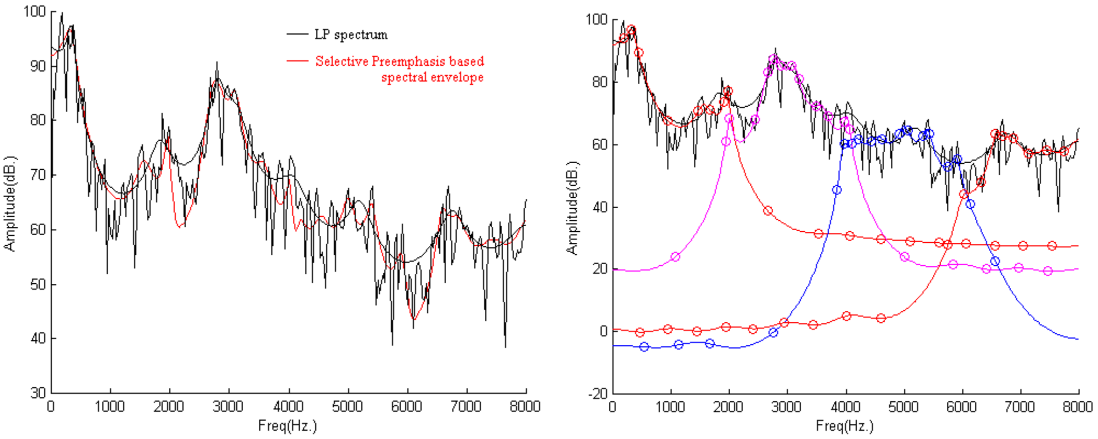


Figure 5.10. LP vs. selective preemphasis based spectral estimation (left). Spectral estimation process from subbands (right) for the phoneme /t/ (female speaker)

In fact, this method is similar to increasing the prediction order in the LP analysis. As 16 KHz is concerned, it is possible to increase the LP order up to 50 and get sufficient spectral detail. However, as we use LSFs for voice conversion and there is a limit for the prediction order for the entire root finding algorithms for calculating LSFs, it becomes impossible to perform detailed spectral analysis. When the sampling rate is 44.1KHz or more, the prediction order should be increased to obtain sufficient spectral detail. Although the procedure described here increases the amount of computation, this can be compensated because if we had the possibility to increase the prediction order beyond the limits, we should also have increased computation time. We can increase the spectral resolution by employing more subbands at a constant prediction order.

It is also possible to use variable prediction orders at each subband and this provides great flexibility in the voice conversion system design. The subbands in which detailed estimation is required can be analyzed using higher prediction orders. The performance of the selective preemphasis based method is evaluated in Chapter 6 in an objective test for spectral modeling accuracy and in subjective tests for voice conversion.

5.2.4. Perceptual Filterbank Design

We have designed a perceptual filterbank for training and transformation using the selective preemphasis system. The center frequencies and bandwidths of the bandpass filters were similar to the filterbank described in Section 4.1.1. The only difference was that each subband had been extended to allow overlap between the neighbouring bands. This method improves the performance at the subband boundaries.

Figure 5.11 shows the magnitude responses of the bandpass filters. The center and cut-off frequencies of the filters are given in Table 5.1. Note that these filters were FIR bandpass filters of order 50 as in Section 4.1.1.

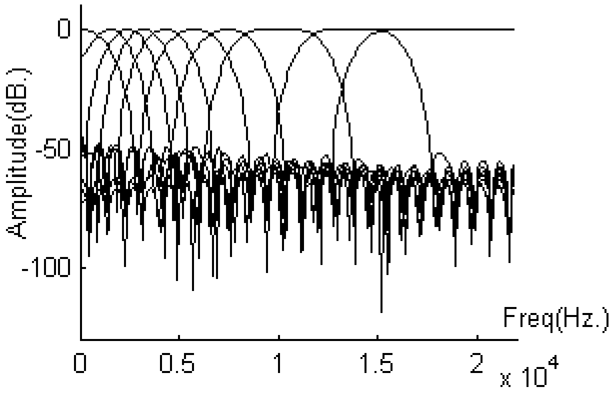
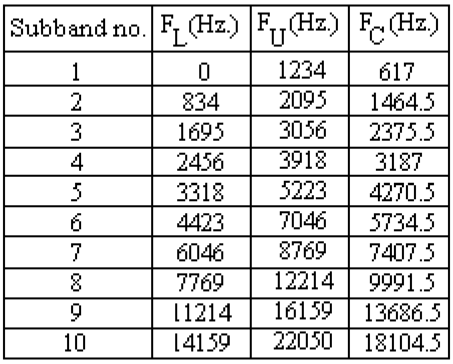


Figure 5.11. Perceptual filterbank for selective preemphasis system



Subband no.	F_L(Hz.)	F_U(Hz.)	F_C(Hz.)
1	0	1234	617
2	834	2095	1464.5
3	1695	3056	2375.5
4	2456	3918	3187
5	3318	5223	4270.5
6	4423	7046	5734.5
7	6046	8769	7407.5
8	7769	12214	9991.5
9	11214	16159	13686.5
10	14159	22050	18104.5

Table 5.1. Cut-off and center frequencies of the bandpass filters

5.2.5. Training

Training is performed by analyzing the source and target utterances using selective preemphasis. LSF vectors for each subband are obtained and the parameters of the first subband are used in the Sentence HMM framework for acoustical alignment. The labels generated using the first subband are used for the remaining subbands. The codebooks are generated for each subband for the source and the target speakers.

An important point in this subband based training approach is the requirement of perfect time synchronization among the analysis of subbands. This can be achieved by either performing analysis using fixed skip rates among frames or employing exactly the same starting and ending instants of the frames for each subband. Using the labels generated for the first subband in the remaining subbands is also necessary for synchronization. The flowchart for the selective preemphasis based training algorithm is shown in Figure 5.12.

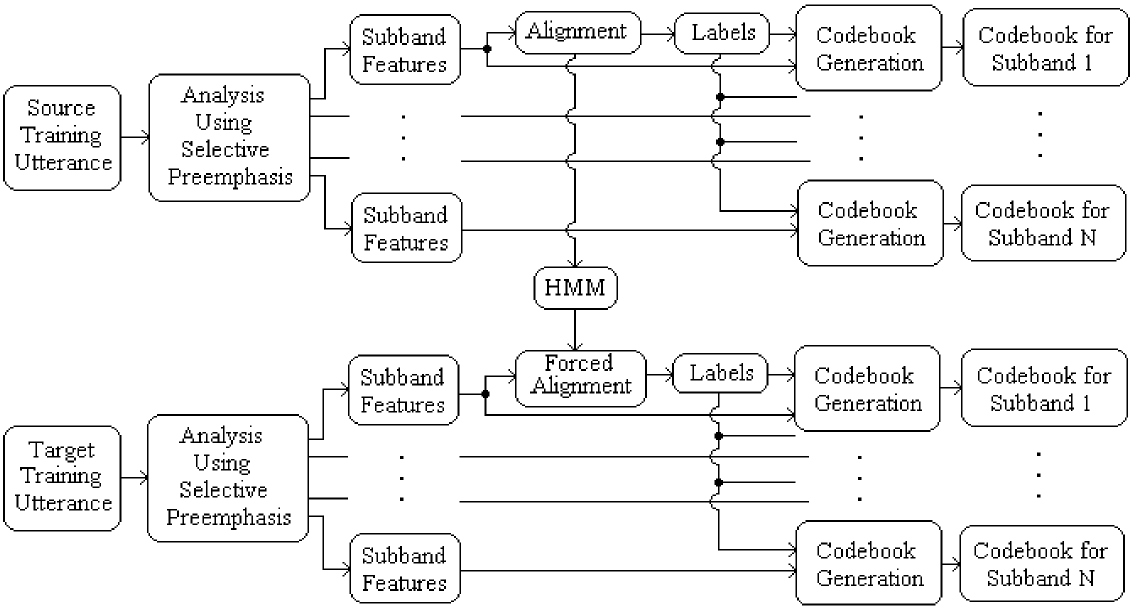


Figure 5.12. Flowchart for selective preemphasis based training algorithm

5.2.6. Transformation

The subband codebooks are used for transforming each subband of the vocal tract spectrum separately. This requires the analysis of the input signal using selective preemphasis. The full-band excitation spectrum is processed separately for pitch scale modifications. Each subband of the vocal tract spectrum is converted employing Equation 3.42 separately for each subband with the corresponding source and target codebook. Note that the closest codebook entries are estimated using the first subband and same indices are used for all subbands. Synthesis is performed using the method described in Section 5.2.2. The output frame spectrum is obtained by multiplying the modified excitation spectrum with the vocal tract spectrum estimated. The flowchart of the transformation system is shown in Figure 5.13.

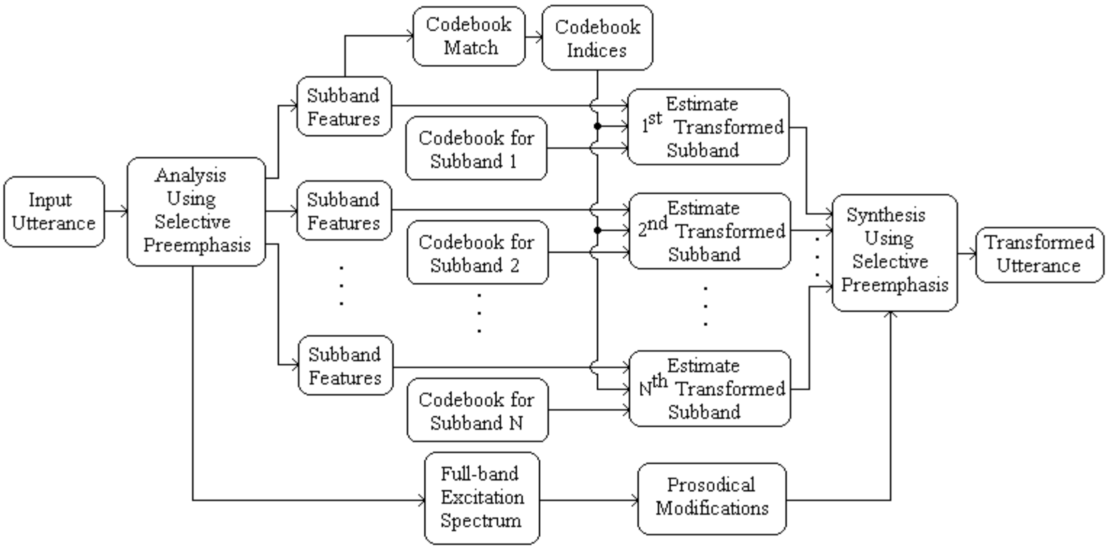


Figure 5.13. Flowchart for selective preemphasis based transformation algorithm

5.3. Comparison of Vocal Tract Transformation Methods

In this section, we present several spectral plots for demonstrating the performance of three different vocal tract transformation methods for different sounds. The methods compared are the full-band approach (Section 3.4), DWT based method (Section 5.1) and selective preemphasis based method (Section 5.2). For Figures 5.14-5.17, we have three types of vocal tract transformation outputs along with the source and target spectral envelope. The original and transformed signals are all from the voice conversion database described in Section 6.1. A prediction order of 50 was used for obtaining the spectral envelopes for a sampling rate of 44.1 KHz.

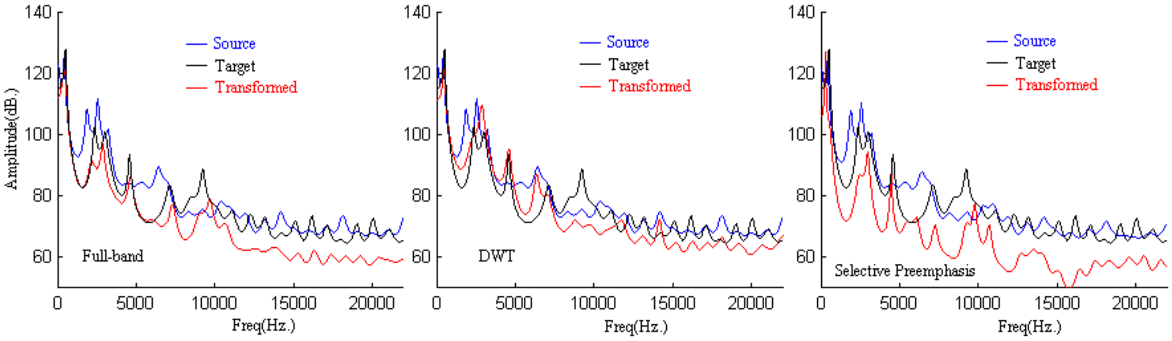


Figure 5.14. Vocal tract transformations for Turkish phoneme /e/ using full-band (left), DWT (middle), and selective preemphasis (right) based methods

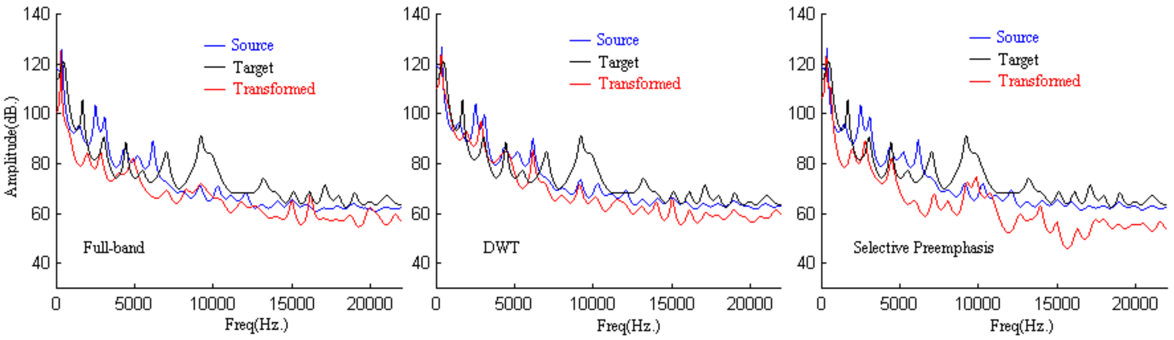


Figure 5.15. Vocal tract transformations for Turkish phoneme /I/ using full-band (left), DWT (middle), and selective preemphasis (right) based methods

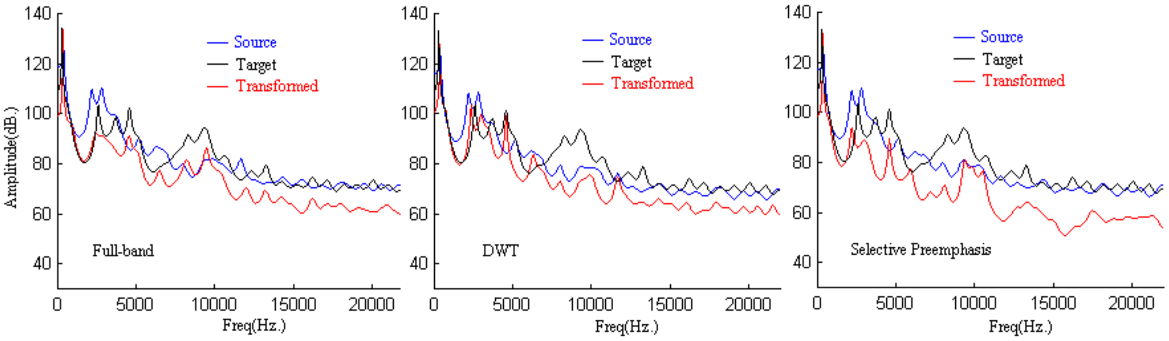


Figure 5.16. Vocal tract transformations for Turkish phoneme /i/ using full-band (left), DWT (middle), and selective preemphasis (right) based methods

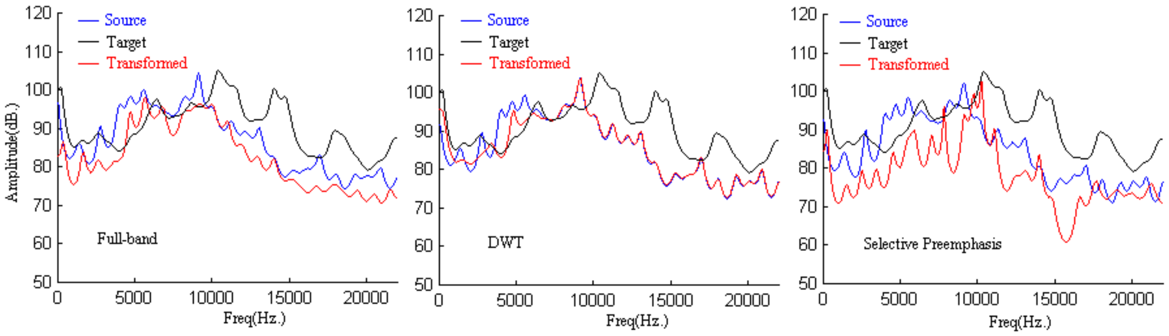


Figure 5.17. Vocal tract transformations for Turkish phoneme /s/ using full-band (left), DWT (middle), and selective preemphasis (right) based methods

5.4. A Segmental Pitch Contour Model for Pitch Contour Transformation

In this section, we describe a new method for pitch contour modeling and transformation. The simplest approach for pitch transformation is to assume the pitch values to be a random variable that is well described by a single Gaussian distribution. In this case, it is fairly easy to estimate and modify the mean and variance of the pitch values between speakers as described in Section 3.6.1. However, the local shapes of the pitch contour segments are not well described and converted using this approach. Although transforming the mean pitch while adjusting the range of pitch values (by adjusting the variance) yields reasonable results, it is clear that the intonational characteristics will be better transformed with a detailed model. For this purpose, we estimate the corresponding pitch contour segments of the source and the target speakers and use this mapping in the transformation stage. If a database containing sufficient intonational information is used, it is possible to generate detailed target pitch contours as we demonstrate below. The method can be described as follows:

- Source and target training utterances are phonetically aligned.
- Pitch contours are extracted & smoothed.
- Target pitch contours are interpolated (linear interpolation or cubic spline interpolation) in the unvoiced parts.
- Voiced segments from the source f0 contours are extracted and for each voiced source f0 segment, the corresponding target segment is found using the phonetic alignment information. We represent the i^{th} source segment as the vector $\mathbf{f0_s^i}$ and the corresponding target segment as $\mathbf{f0_t^i}$.
- The vectors $\mathbf{f0_s^i}$'s and $\mathbf{f0_t^i}$'s are written into a pitch contour codebook file along with the mean and the standard deviation of the f0 values of the source and target speakers. This completes the training procedure for pitch transformations.
- In the transformation stage, the voiced segments, $\mathbf{f0^j}$'s, of the input pitch contour are found. We denote the length of segment as N^j. Source and target codebook entries are interpolated to the length N^j before calculating the distance measure d^i. Then, the Euclidean distance of the current input segment, $\mathbf{f0^j}$, to each source codebook entry is estimated (Equation 5.4). The DC shifts of the segments are removed before distance calculation in order to ensure that they do not cause an

increase in the distance when two segments were similar in shape but differing in mean value.

$$d^i = \sum_{n=0}^{N^j-1} \mid (f0^j(n) - \mu^j) - (f0^i_s(n) - \mu^i_s) \mid^2 \tag{5.4}$$

where μ_j denotes the mean of the vector $\mathbf{f0^j}$ and μ^i_s is the mean of the interpolated version of the vector $\mathbf{f0_s^i}$.

- The distances are normalized such that they sum up to unity using Equation 5.5.

$$d^i_{normalized} = \frac{d^i}{\sum_{all\ i} d^i} \tag{5.5}$$

- A weight for each source segment in the codebook is estimated using the normalized distances (d^i's) employing Equation 5.6. The new weights are normalized to sum up to unity again. We have used α=500 as the weighting factor. This ensures that only a few close matches from the codebook are included in the generation of the synthetic codebook. For smaller values of α, a weighted average of more and more codebook entries will be used which will result in a smoother pitch contour.

$$w^i = \exp(-\alpha d^i_{normalized}) \tag{5.6}$$

$$w^i_{normalized} = \frac{w^i}{\sum_{all\ i} w^i} \tag{5.7}$$

- All target codebook segments are interpolated to match the length $\mathbf{f0^j}$ and the synthetic pitch contour segment is estimated using the weights and the target codebook entries:

$$f0^j_{synth} = \sum_{n=0}^{N^j-1} w^i_{normalized} f0^i_t(n) \tag{5.8}$$

where $\mathbf{f0^j_{synth}}$ is the pitch contour segment vector to be used in pitch scaling.

The procedure above is performed in the training stage. However, it is also possible to use a separate database for modeling the source and target pitch contours. Figure 5.18 shows two examples of modeling pitch contours by this method. Note that, the model tracks the general shape of the pitch contours well but there are mismatches when the contours possess sudden jumps. These jumps may be both due to the intonational characteristics of the speakers and pitch estimation errors. So, employing an intonationally rich pitch contour database extracted with a robust pitch detection algorithm is required. We have used both autocorrelation (Rabiner and Schafer, 1978) and RAPT (Talkin, 1995) algorithms for pitch detection.

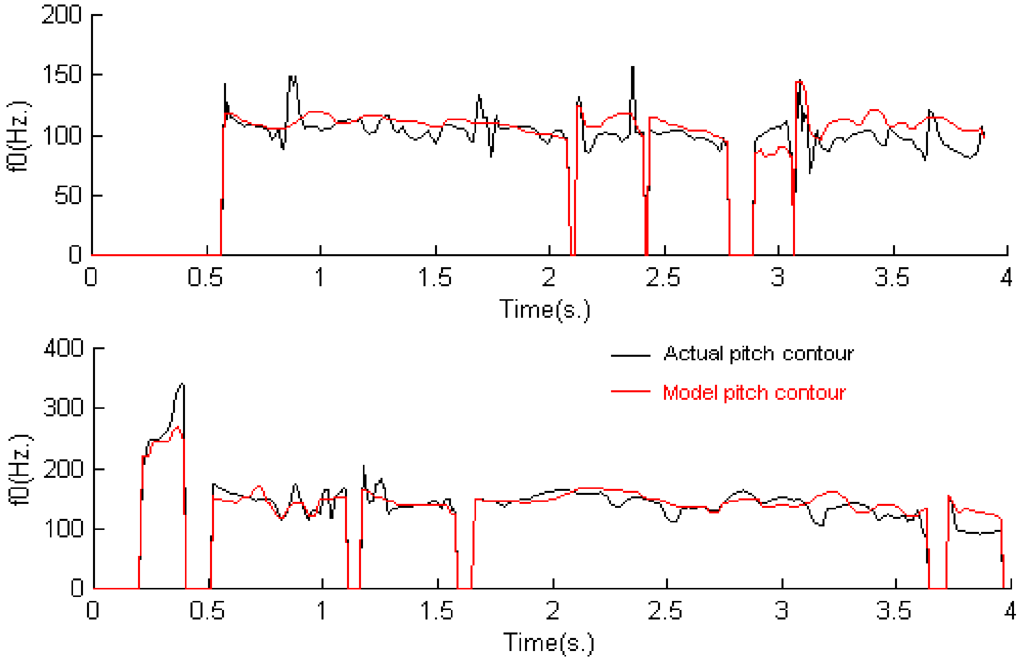


Figure 5.18. Segmental pitch contour model for the utterance of a male speaker (top), and a female speaker (bottom)

In Figure 5.18, the original pitch contours and the estimated pitch contours using the segmental model for an utterance of a male speaker (top) and a female speaker (bottom) is shown. The black contours are original contours from Turkish sentences that were not used for generating the pitch contour codebooks. Red contours are estimated using the segmental pitch codebooks.

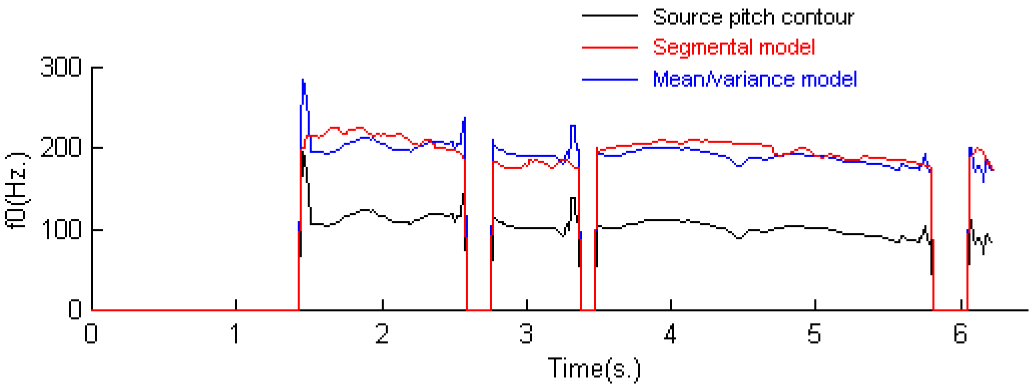


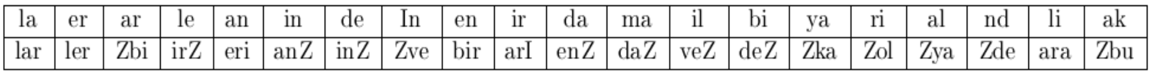
Figure 5.19. Pitch contour transformation with the segmental pitch model. The source contour is from a male utterance and the target is the same utterance by a female speaker

In Figure 5.19, pitch contour transformation using the segmental pitch contour model is demonstrated. The blue contour corresponds to the output of the mean/variance model for the same source sentence for comparison. Notice that the shape of the source pitch contour (black) is also modified in the output contour produced by the segmental pitch model (red) while increasing the pitch. In Chapter 6, we use this method in voice conversion and compare the subjective performance of the model with the simple mean/variance model which was described in Section 3.6.1.

6. TEST DESIGN AND EVALUATIONS

6.1. Design of the Voice Conversion Database

We have collected several databases both in Turkish and in English during this study. The most general databasecontains 18 Turkish speakers (10 male, 8 female) in different ages. It was used in evaluating the new methods described in Chapter 5, and for acoustic feature transplantations. It consists of 31 sentences (containing 3-5 words) and 50 words for training. Five long sentences (containing 6-8 words) and 10 words were recorded for testing. All recordings were at 48 KHz and in 16-bit PCM files. We have considered the bigram and trigram probabilities in Turkish in designing this database. Special care was taken for including the most common bigrams and trigrams. We have used a simple text search algorithm with a Turkish text database of approximately 220000 words to estimate the probabilities. The results described in (Yapanel, 2000) for Turkish trigram probabilities were also used as they were obtained from a larger corpus of 2.2 million words. The most common 20 bigrams and trigrams are given in Table 6.1.



la	er	ar	le	an	in	de	In	en	ir	da	ma	il	bi	ya	ri	al	nd	li	ak
lar	ler	Zbi	irZ	eri	anZ	inZ	Zve	bir	arI	enZ	daZ	veZ	deZ	Zka	Zol	Zya	Zde	ara	Zbu

Table 6.1. The most common 20 bigrams and trigrams of Turkish

In our previous studies, we have used other voice conversion databases that were more restricted in terms of corpus size and number of speakers as compared to the general database described above (Turk and Arslan, 2002, and Ormancı *et al.*, 2002). In the subjective tests performed in Section 4.1, the database was restricted to 8 Turkish speakers (four male, four female) and contained 15 sentences and 20 words recorded at 44.1 KHz. Another database was used to compare the DWT based vocal tract conversion system with the full-band system as described in Section 6.2.1 and it contained 10 speakers (five male, five female). In this case, one female and three males were native American-English speakers. More information regarding the databases is included in the description of the tests.

6.2. Design of Perceptual Tests for Voice Conversion

In the following sub-sections, we describe several subjective tests for the evaluation of the new methods proposed.

6.2.1. Comparison of the DWT Based System with the Full-band System

In (Turk and Arslan, 2002), we have performed ABX listening tests to evaluate the subjective performance of the DWT based vocal tract conversion method described in Section 5.1. We have used different combinations of five male and five female speakers as the source and the target. Four speakers (one female and three males) were native American-English speakers and the remaining were native Turkish speakers. Four types of voice conversion is performed as far as the gender of the source and the target is concerned (female-to-female, female-to-male, male-to-female,and male-to-male). First, training and test utterances were recorded at 44.1KHz. Full-band and subband based codebooks are generated by two separate training sessions for each conversion. We have not used preemphasis for the full-band case. A linear prediction order of 50 was used for full-band based training and transformation. In the subband case, four subbands were employed and only the first subband is converted using an LP order of 18. Each of the 20 subjects listened to 10 sentences and 10 words that were transformed using the subband and full-band codebooks. The subjects were provided with three recordings each time they were asked to make a decision: (A) Full-band based conversion output, (B) Subband based conversion output, and (X) Target recording. The conversion outputs were presented in random order and the listener was asked to judge whether (A) or (B) sounds more like the target speech (X). The order of the full-band and the subband based output was also changed randomly. The subband based voice conversion output was preferred over the full-band based output by 92.9 percent for sentences, 91.3 percent for words, resulting in an overall preference rate of 92.1 percent.

6.2.2. Evaluation of New Methods

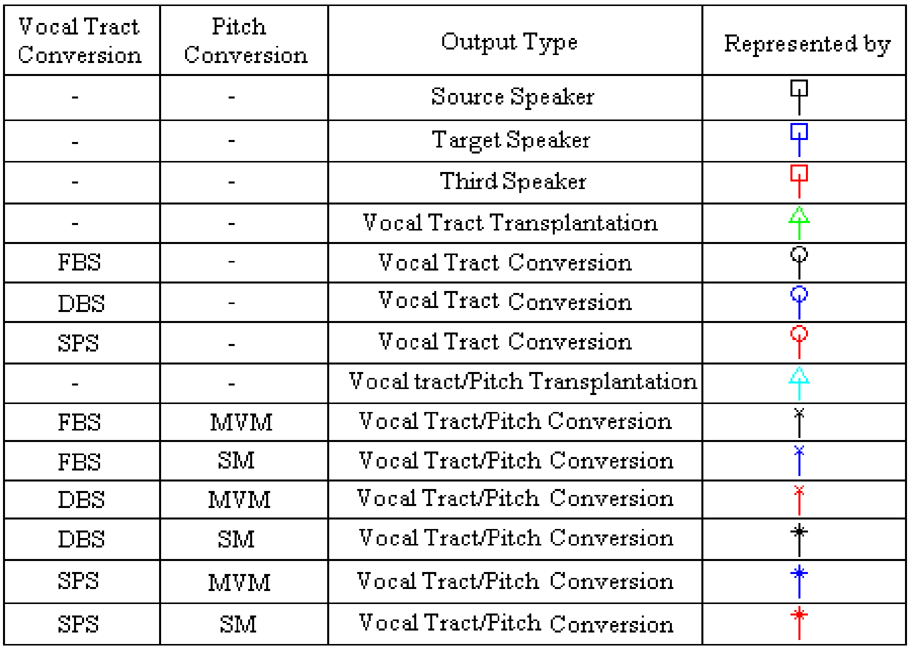
In this section, we describe a subjective testing procedure for assessing the performance of the proposed methods for voice conversion regarding the similarity of the output to target speaker's voice and its quality. All combinations of three vocal tract and two pitch conversion methods are evaluated. The vocal tract conversion methods are the full-band system in STASC with preemphasis (Section 3.4), the DWT based system (Section 5.1), and the selective preemphasis system (Section 5.2). For pitch conversion, we have employed the mean/variance model (Section 3.6.1) and the segmental pitch contour model (Section 5.4). We use the following short-hand notations for these methods:

- Full-band system: FBS
- DWT Based System: DBS
- Selective Preemphasis System: SPS
- Mean/Variance Model: MVM
- Segmental Model: SM

We use the same shorthand notations for the vocal tract and vocal tract/pitch contour transplantations as in Section 4.2:

- Vocal tract transplantation: VTT
- Vocal tract/Pitch contour transplantation: VTT-PCT

A subset of the general database was used for evaluations including four male and four female speakers. Four types of conversion is performed as the gender of the source and the target speakers are considered: M→M, M→F, F→M, and F→F where M denotes a male speaker and F denotes a female speaker. We have used 30 sentences and 50 words by the source and the target speakers in training and generated separate codebooks for the sentences and the words. Five sentences and 10 words are converted with different methods using corresponding codebooks (i.e. word codebooks were used for converting words and vice versa). Considering all the combinations that employ vocal tract conversion, we have the cases shown in Table 6.2.



Vocal Tract Conversion	Pitch Conversion	Output Type	Represented by
-	-	Source Speaker	□
-	-	Target Speaker	□
-	-	Third Speaker	□
-	-	Vocal Tract Transplantation	△
FBS	-	Vocal Tract Conversion	○
DBS	-	Vocal Tract Conversion	○
SPS	-	Vocal Tract Conversion	○
-	-	Vocal tract/Pitch Transplantation	△
FBS	MVM	Vocal Tract/Pitch Conversion	×
FBS	SM	Vocal Tract/Pitch Conversion	×
DBS	MVM	Vocal Tract/Pitch Conversion	×
DBS	SM	Vocal Tract/Pitch Conversion	*
SPS	MVM	Vocal Tract/Pitch Conversion	*
SPS	SM	Vocal Tract/Pitch Conversion	*

Table 6.2. Voice conversion methods tested

The test database is prepared as follows:

- The original recordings are selected from the database (at 48KHz)
- Downsampling to 44KHz
- Automated phonetic labeling followed by manual correction
- For each source/target pair, we have trained the fullband, DWT and selective preemphasis systems for vocal tract conversion. The segmental pitch contour model was trained while performing fullband training – i.e. we have extracted the pitch contours of the training utterances and used them to estimate the parameters of the segmental model.
- Nine types of conversion methods are employed as described in Table 6.2.

In the tests, 10 subjects were provided with several utterance triples. The first and the second utterances in each triple are the original recordings of the source and target speakers. The third utterance contains the output to be evaluated by the subject. It is one of the following:

- An original recording (of the source, target, or a third speaker as given in rows 1-3 of Table 6.2),

- A transplantation output (either vocal tract only or vocal tract/pitch transplantation as given in row 4 or row 8 of Table 6.2),
- A conversion output (obtained using one of the voice conversion methods given in rows 5-7 and rows 9-14 of Table 6.2).

The subjects have provided three scores:

- Identity Score: For this score, the subjects have provided the decisions “source”, “target”, and “in between”. These decisions are mapped onto a numerical scale by assigning “0.0” for the “source” decisions, “0.5” for “in between” decisions, and “1.0” for the target decisions. The subjects were told to respond with an “in between” decision with a low confidence score when the output sounds like a third speaker. This does not result in loss of generality because if the subjects think that the output sounds like both speakers, they should provide a higher confidence score. It is possible to discriminate between “in between” and “third speaker” cases using the confidence score.
- Confidence Score: This score indicates the confidence of the subject on the identity score. It is in the range 1-5, where 1 corresponds to low confidence and 5 corresponds to high confidence. For the case that the output sounds like a third speaker, the speakers were told to respond with the lowest confidence score as explained above.
- Quality Score: This score is a measure of the quality of the output as compared to the first two files. This score is also in the range 1-5. 1 corresponds to very low quality. A quality score of 5 indicates that the quality of the output is similar to the original recordings.

The subjects have listened to 112 triples (72 voice conversion outputs, 16 transplantation outputs, 8 source recordings, 8 target recordings, 8 recordings of a third speaker). The results are evaluated in a similar manner as in the transplantation tests (Section 4.2). The only difference is that the subjects have also provided a score on the quality of the output. Note that we have normalized all scores to the range [0.0, 1.0]. The test took 20-30 minutes for each subject. In Figures 6.1-6.6, we present the subjective test results. In general, we observe that the source, target and third speakers were identified correctly. The identity score for the source speaker is low indicating that average similarity to target speaker is less. The target speaker had high identity scores (close to 1.0). The subjects have responded with identity scores close to 0.5 with low confidence scores in the case of "third speaker". This is expected since the subjects were told to respond with low confidence scores when the output sounded like a third speaker. The quality scores for all the original recordings (source, target, and third speaker) were close to 1.0 as expected. Figures 6.1 and 6.2 show the statistics for all types of utterances (i.e the words and the sentences together). Figures 6.3 and 6.4 are for words. Finally, Figures 6.5 and 6.6 are for the sentences. As the overall case is considered, we observe that converting only the vocal tract does not produce convincing results. Even the vocal tract transplantation case was evaluated as in between the source and the target speaker. Vocal tract conversions generally had higher identity scores when the source is a male speaker. The performance is reduced when the source is a female speaker. All voice conversion methods produce more convincing results in terms of similarity to the target speaker when pitch transformation strategies are involved as expected. However, we observe a reduction in both the confidence and the quality scores as the amount of processing increases. Vocal tract only transplantations and conversions received higher scores in terms of confidence and quality in general. We observe different tendencies as the gender of the source and target speaker pairs change for different vocal tract conversion methods. The fullband based method is more robust in different gender combinations. The segmental pitch model improves the identity scores in general.

Tables 6.3-6.5 show the interquartile ranges of the scores. We observe that the lowest ranges were for the original recordings. Vocal tract only conversions had lower interquartile ranges in the case of identity scores. For the confidence scores, the ranges were lower for the new vocal tract conversion methods when they were used with pitch transformation methods. The interquartile ranges for the quality scores were similar for all type of conversions and transplantations. We observe that the performance of the full-band based method is improved when preemphasis was employed. The full-band based method has performed better as the results are compared with the results of Section 6.2.1.

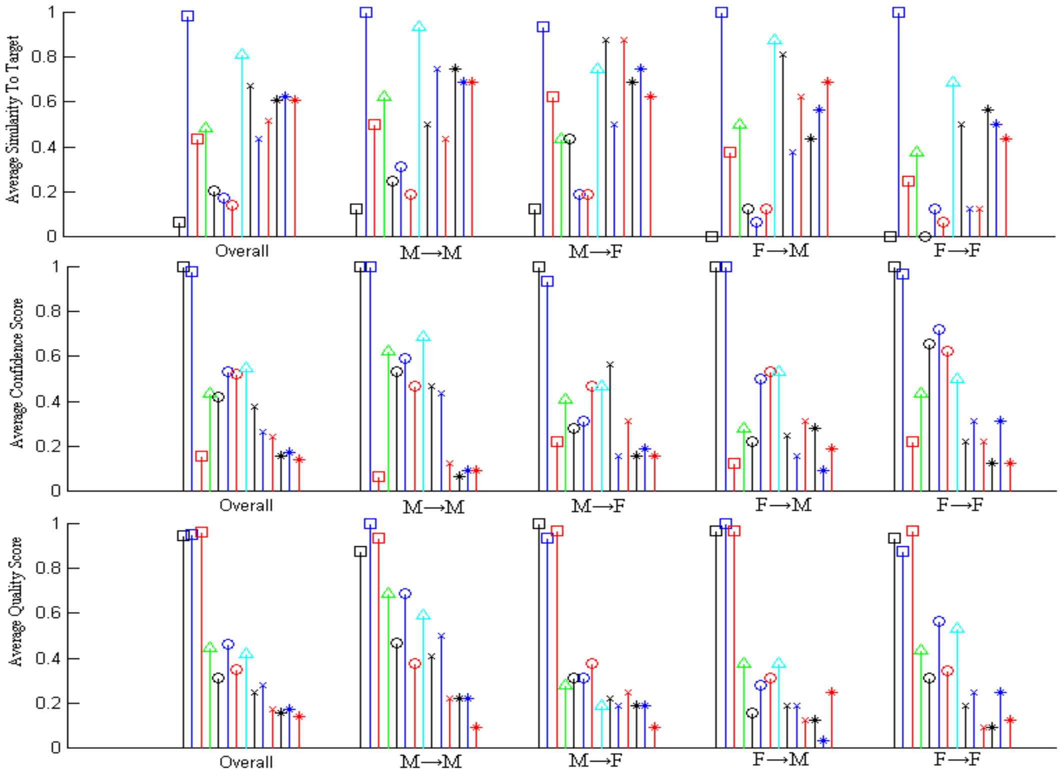


Figure 6.1. Voice conversion subjective test results for all utterances

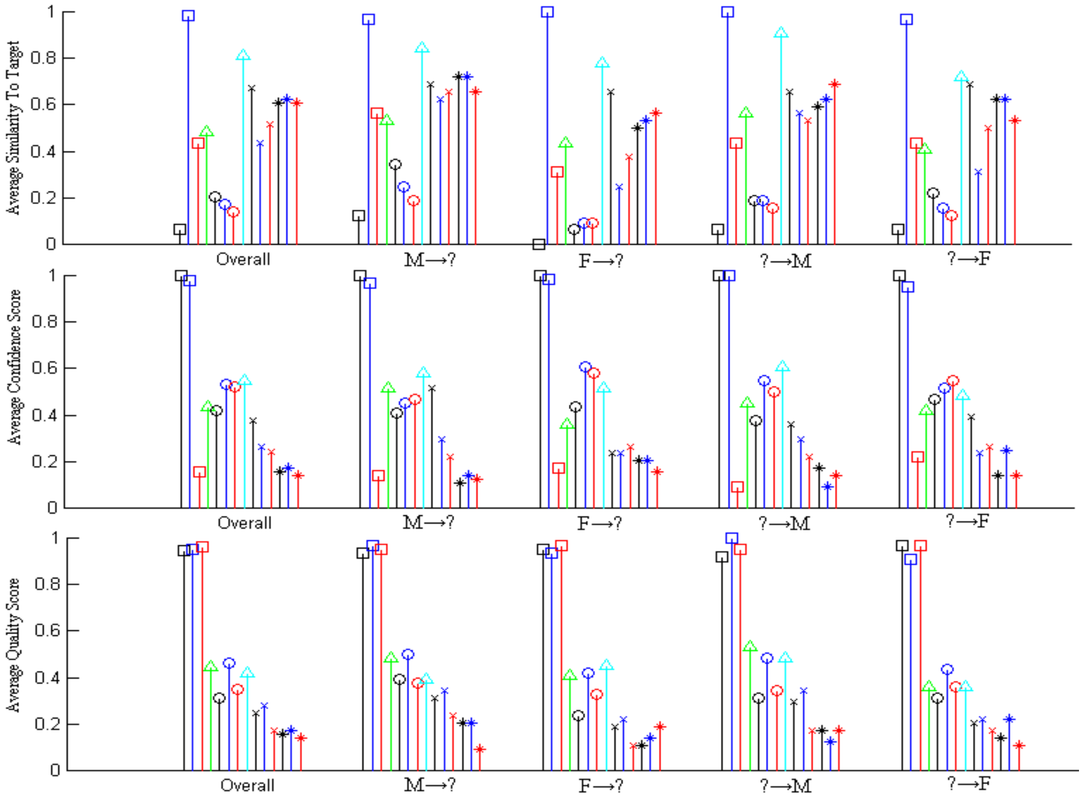


Figure 6.2. Voice conversion subjective test results for all utterances

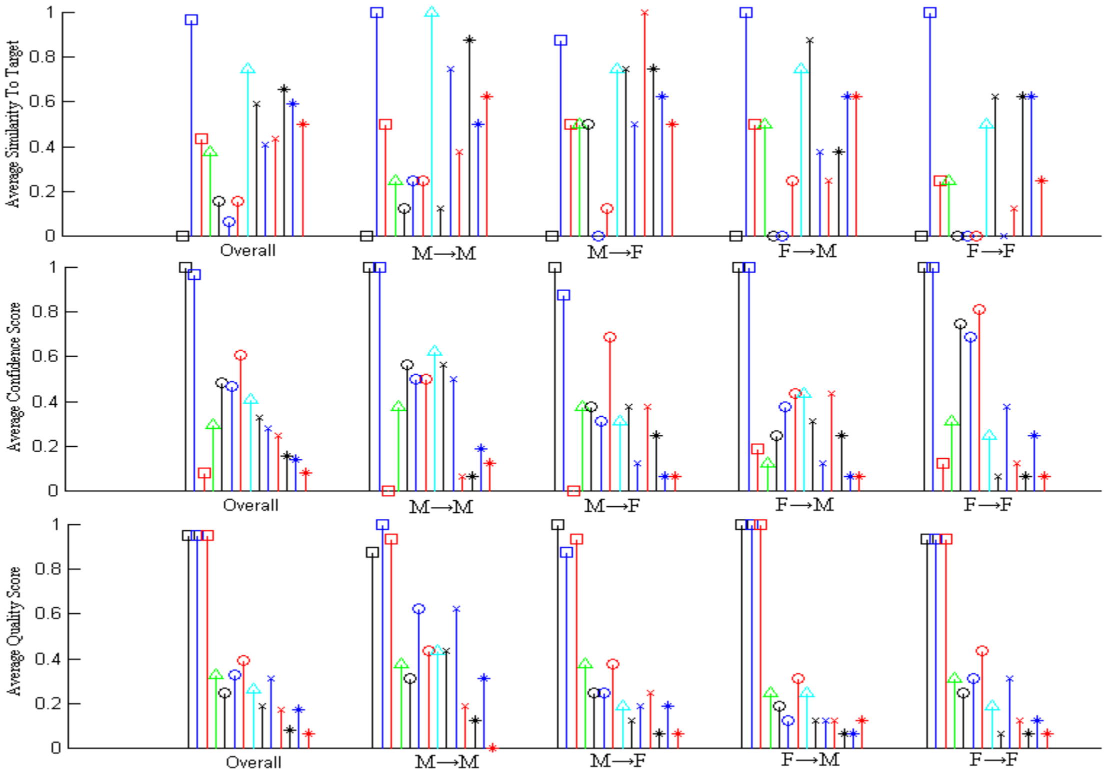


Figure 6.3. Voice conversion subjective test results for words

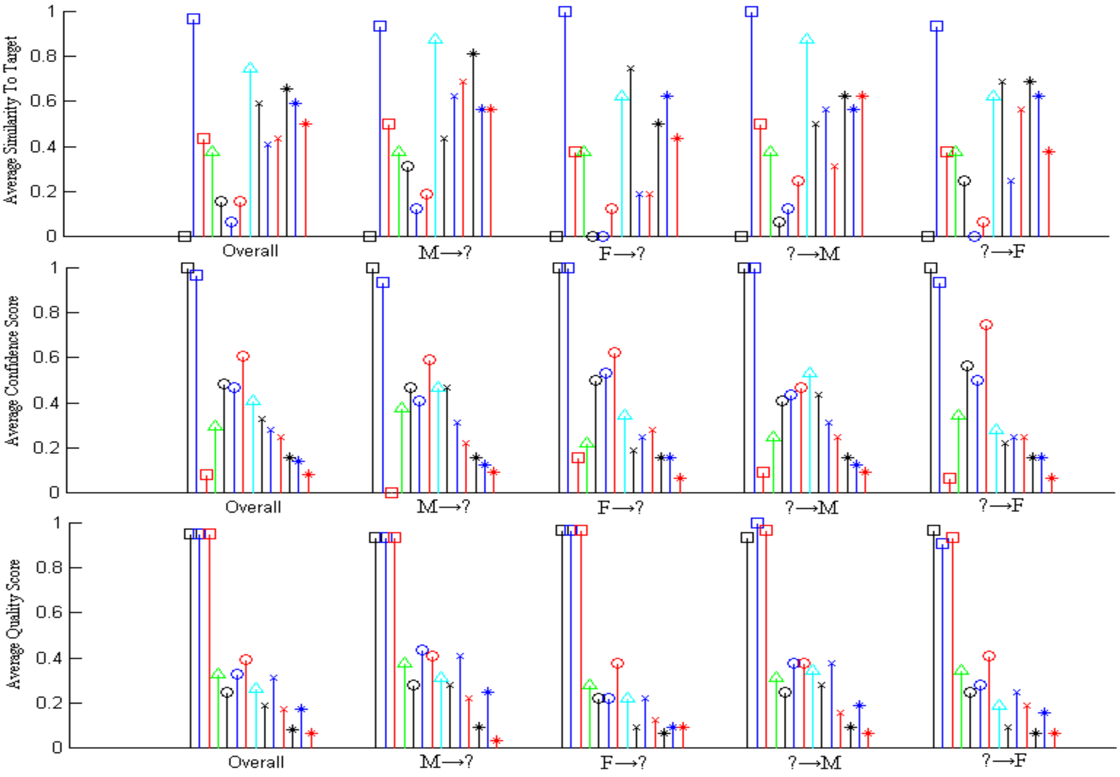


Figure 6.4. Voice conversion subjective test results for words

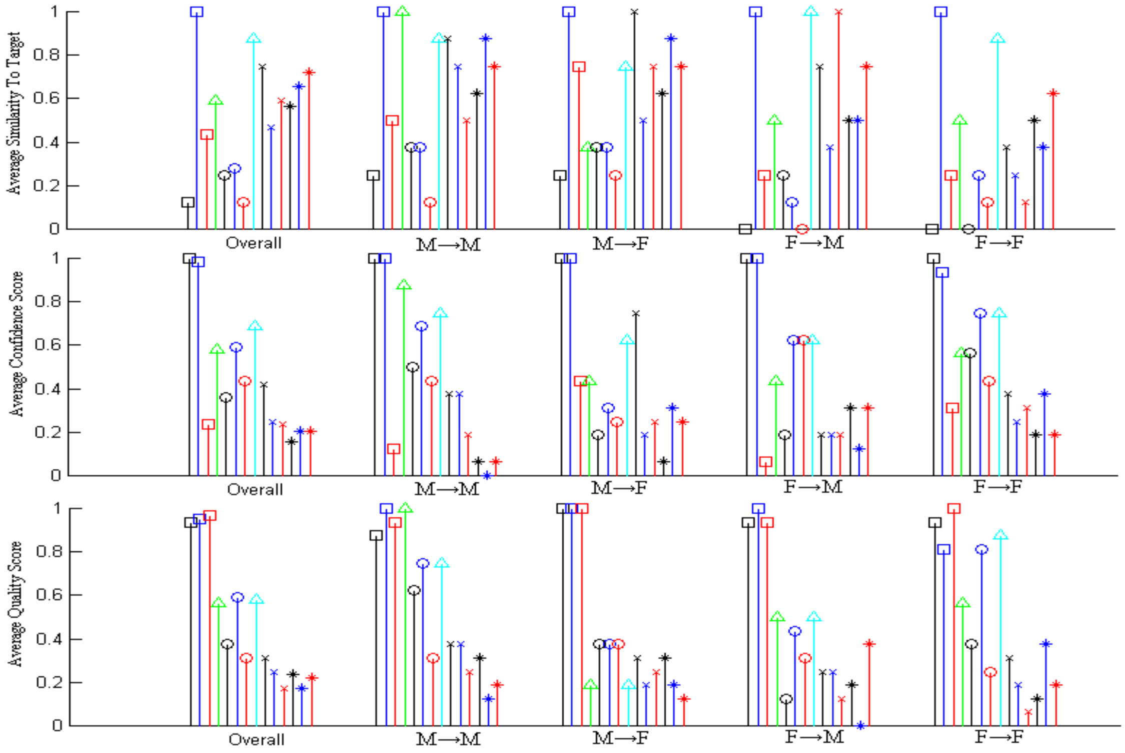


Figure 6.5. Voice conversion subjective test results for sentences

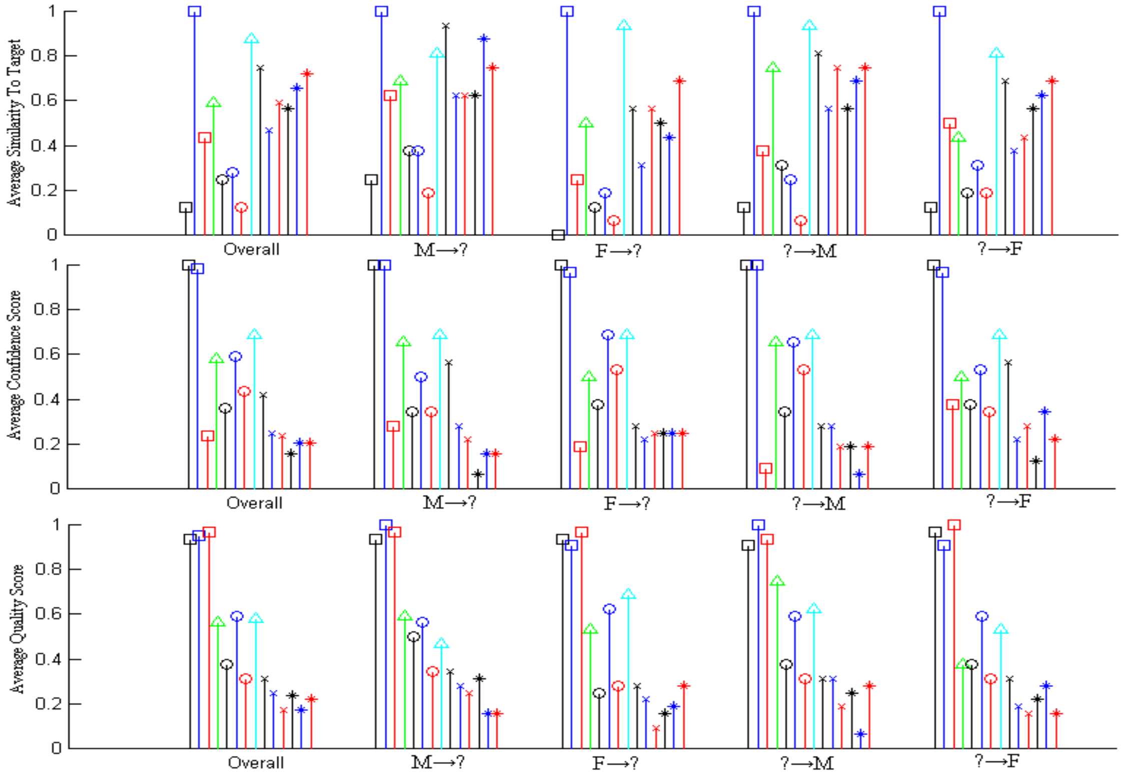
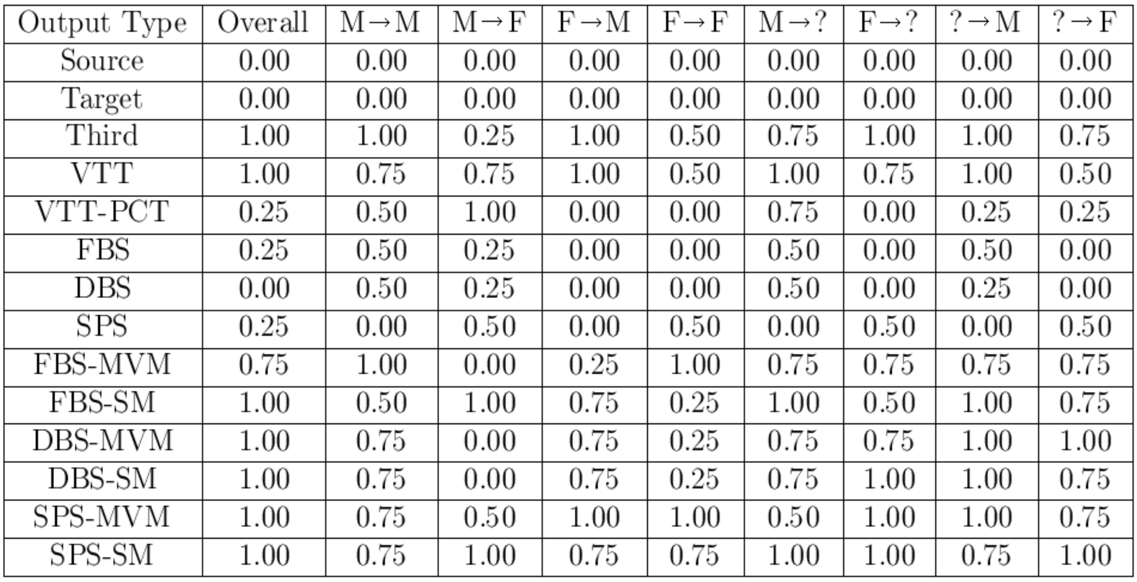
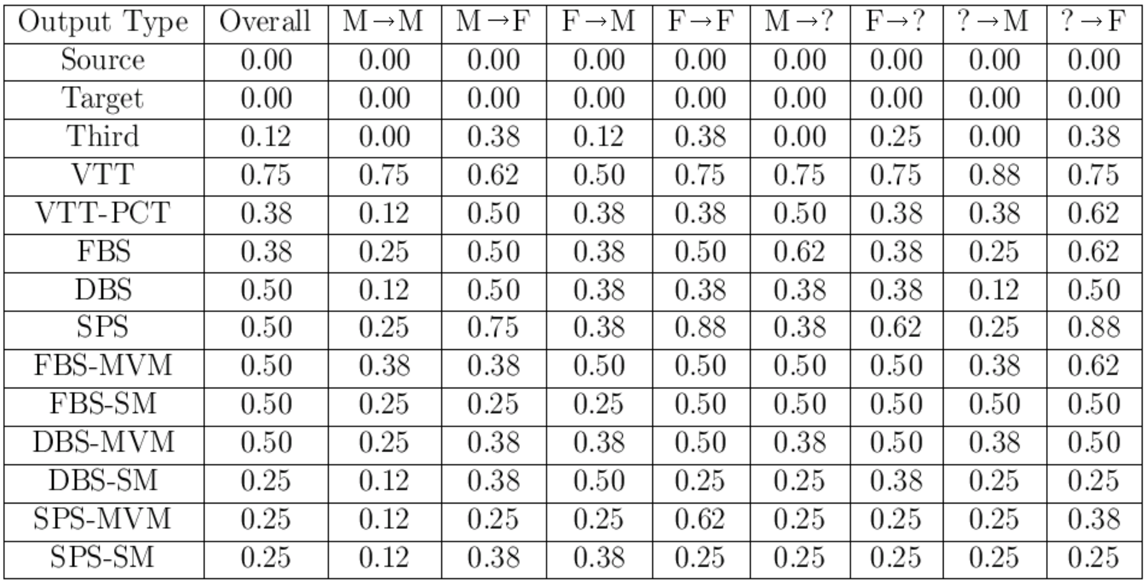


Figure 6.6. Voice conversion subjective test results for sentences



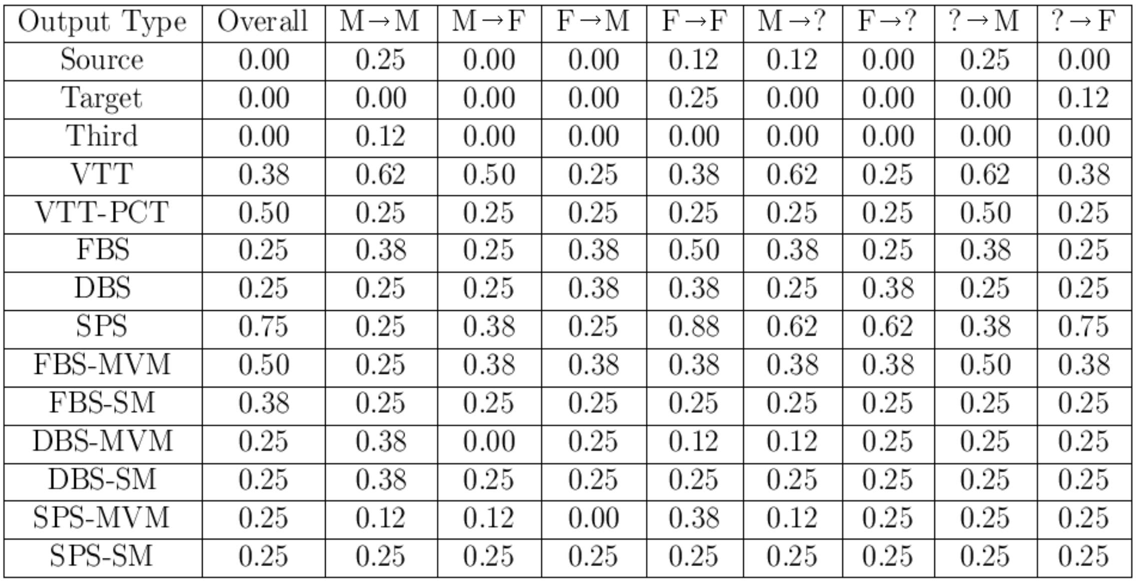
Output Type	Overall	M→M	M→F	F→M	F→F	M→?	F→?	?→M	?→F
Source	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Target	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Third	1.00	1.00	0.25	1.00	0.50	0.75	1.00	1.00	0.75
VTT	1.00	0.75	0.75	1.00	0.50	1.00	0.75	1.00	0.50
VTT-PCT	0.25	0.50	1.00	0.00	0.00	0.75	0.00	0.25	0.25
FBS	0.25	0.50	0.25	0.00	0.00	0.50	0.00	0.50	0.00
DBS	0.00	0.50	0.25	0.00	0.00	0.50	0.00	0.25	0.00
SPS	0.25	0.00	0.50	0.00	0.50	0.00	0.50	0.00	0.50
FBS-MVM	0.75	1.00	0.00	0.25	1.00	0.75	0.75	0.75	0.75
FBS-SM	1.00	0.50	1.00	0.75	0.25	1.00	0.50	1.00	0.75
DBS-MVM	1.00	0.75	0.00	0.75	0.25	0.75	0.75	1.00	1.00
DBS-SM	1.00	0.75	0.00	0.75	0.25	0.75	1.00	1.00	0.75
SPS-MVM	1.00	0.75	0.50	1.00	1.00	0.50	1.00	1.00	0.75
SPS-SM	1.00	0.75	1.00	0.75	0.75	1.00	1.00	0.75	1.00

Table 6.3. Interquartile ranges for the identity scores



Output Type	Overall	M→M	M→F	F→M	F→F	M→?	F→?	?→M	?→F
Source	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Target	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Third	0.12	0.00	0.38	0.12	0.38	0.00	0.25	0.00	0.38
VTT	0.75	0.75	0.62	0.50	0.75	0.75	0.75	0.88	0.75
VTT-PCT	0.38	0.12	0.50	0.38	0.38	0.50	0.38	0.38	0.62
FBS	0.38	0.25	0.50	0.38	0.50	0.62	0.38	0.25	0.62
DBS	0.50	0.12	0.50	0.38	0.38	0.38	0.38	0.12	0.50
SPS	0.50	0.25	0.75	0.38	0.88	0.38	0.62	0.25	0.88
FBS-MVM	0.50	0.38	0.38	0.50	0.50	0.50	0.50	0.38	0.62
FBS-SM	0.50	0.25	0.25	0.25	0.50	0.50	0.50	0.50	0.50
DBS-MVM	0.50	0.25	0.38	0.38	0.50	0.38	0.50	0.38	0.50
DBS-SM	0.25	0.12	0.38	0.50	0.25	0.25	0.38	0.25	0.25
SPS-MVM	0.25	0.12	0.25	0.25	0.62	0.25	0.25	0.25	0.38
SPS-SM	0.25	0.12	0.38	0.38	0.25	0.25	0.25	0.25	0.25

Table 6.4. Interquartile ranges for the confidence scores



Output Type	Overall	M→M	M→F	F→M	F→F	M→?	F→?	?→M	?→F
Source	0.00	0.25	0.00	0.00	0.12	0.12	0.00	0.25	0.00
Target	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.12
Third	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VTT	0.38	0.62	0.50	0.25	0.38	0.62	0.25	0.62	0.38
VTT-PCT	0.50	0.25	0.25	0.25	0.25	0.25	0.25	0.50	0.25
FBS	0.25	0.38	0.25	0.38	0.50	0.38	0.25	0.38	0.25
DBS	0.25	0.25	0.25	0.38	0.38	0.25	0.38	0.25	0.25
SPS	0.75	0.25	0.38	0.25	0.88	0.62	0.62	0.38	0.75
FBS-MVM	0.50	0.25	0.38	0.38	0.38	0.38	0.38	0.50	0.38
FBS-SM	0.38	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
DBS-MVM	0.25	0.38	0.00	0.25	0.12	0.12	0.25	0.25	0.25
DBS-SM	0.25	0.38	0.25	0.25	0.25	0.25	0.25	0.25	0.25
SPS-MVM	0.25	0.12	0.12	0.00	0.38	0.12	0.25	0.25	0.25
SPS-SM	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25

Table 6.5. Interquartile ranges for the quality scores

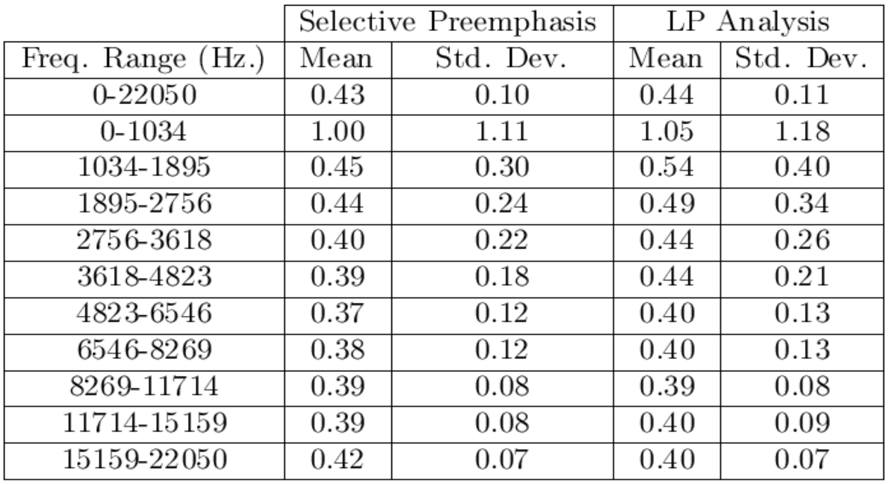
6.3. Objective Tests

6.3.1. Comparison of LP and Selective Preemphasis Based Spectral Estimation

We have performed an objective test for the comparison of the spectral estimation performance of LP analysis and selective preemphasis system. For this purpose, recordings from different speakers were analyzed. We have calculated the spectral distance measure defined by Equation 6.1 frame-by-frame for both methods.

$$D_n^i = \frac{1}{\omega_2 - \omega_1}\int_{\omega_1}^{\omega_2} (10\log_{10}(P_{orig}(\omega)) - 10\log_{10}(P_{est}(\omega)))^2\, d\omega \qquad (6.1)$$

A fixed frame duration of 30 msec. with 10 msec. skip rate is used. The recordings were at 44.1 KHz. The prediction order was 50 for the LP analysis. We have used the analysis scheme described in Section 5.2 with the bandpass filterbank given in Figure 5.11 for selective preemphasis based spectral estimation. The prediction order was 24 for all subbands. In Table 6.6, the mean and the standard deviations of the distances obtained are given. The selective preemphasis based system performs better than LP analysis in terms of spectral distance at a lower prediction order.



	Selective Preemphasis		LP Analysis	
Freq. Range (Hz.)	Mean	Std. Dev.	Mean	Std. Dev.
0-22050	0.43	0.10	0.44	0.11
0-1034	1.00	1.11	1.05	1.18
1034-1895	0.45	0.30	0.54	0.40
1895-2756	0.44	0.24	0.49	0.34
2756-3618	0.40	0.22	0.44	0.26
3618-4823	0.39	0.18	0.44	0.21
4823-6546	0.37	0.12	0.40	0.13
6546-8269	0.38	0.12	0.40	0.13
8269-11714	0.39	0.08	0.39	0.08
11714-15159	0.39	0.08	0.40	0.09
15159-22050	0.42	0.07	0.40	0.07

Table 6.6. Mean and standard deviations of spectral distances (dB/Hz) using LP analysis and selective preemphasis based analysis

6.3.2. Source to Target and Transformed to Target Distances

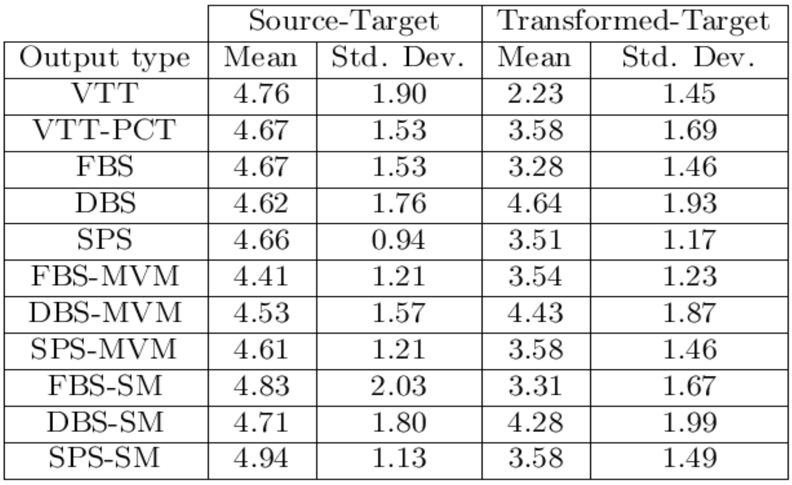
Although subjective testing provides extensive information on the performance of the new voice conversion methods, it is possible to employ objective distance measures for a numerical evaluation. The procedure is as follows:

- The source, target and transformed utterances are labeled. It is possible to use manual labels or automatically generated labels by the Sentence HMM method.
- The average LSF distance (L) is estimated for each phoneme in the case for manual labeling, or state in the case of a Sentence HMM. Equations 6.3 and 6.4 provide a measure based on LSF distances. They can be employed to calculate a measure of similarity of two spectra. In fact these are the same equations used in STASC for vocal tract transformation as described in Section 3.4.2 (Equations 3.29 and 3.30).
- The overall statistics (mean and standard deviations) are estimated using the source-target-transformed triples generated for the subjective tests in Section 6.2.2.

$$L = \sum_{k=1}^{P} h_k \; |lsf_k^1 - lsf_k^2| \tag{6.3}$$

$$h_k = \frac{1}{argmin(|\; lsf_k - lsf_{k-1} \;|, |\; lsf_k - lsf_{k+1} \;|)} \quad for \quad k = 1, \ldots, P \tag{6.4}$$

We have used the vocal tract and vocal tract/pitch conversion outputs along with the corresponding source and target pairs to obtain the results given in Table 6.7. We have also included the transplantation outputs. In particular, we expect to reduce the spectral distance between the source and target utterances by voice conversion. So the transformed to target spectral distances should be lower than the source to target distances in general.



	Source-Target		Transformed-Target	
Output type	Mean	Std. Dev.	Mean	Std. Dev.
VTT	4.76	1.90	2.23	1.45
VTT-PCT	4.67	1.53	3.58	1.69
FBS	4.67	1.53	3.28	1.46
DBS	4.62	1.76	4.64	1.93
SPS	4.66	0.94	3.51	1.17
FBS-MVM	4.41	1.21	3.54	1.23
DBS-MVM	4.53	1.57	4.43	1.87
SPS-MVM	4.61	1.21	3.58	1.46
FBS-SM	4.83	2.03	3.31	1.67
DBS-SM	4.71	1.80	4.28	1.99
SPS-SM	4.94	1.13	3.58	1.49

Table 6.7. Mean and standard deviations of source to target and transformed to target LSF distances

7. CONCLUSIONS

In the first part of this study, we have investigated the role of different factors on perception of speaker identity (Chapter 4). We have found that the frequency range [1034 Hz, 1895 Hz] is the most important spectral region. The frequency range between [1895 Hz, 2756 Hz] is of second importance. We have designed another subjective test for evaluating the relevance of different acoustic features in the perception of speaker identity. Four acoustic features were transplanted using PSOLA based methods: Vocal tract, pitch contour, phonemic durations, and energy contour. The vocal tract/pitch/duration transplantation had the highest scores in terms of similarity to the target speaker. In the case of single feature transplantations, vocal tract was the most important feature. Pitch and duration had less importance than the vocal tract. Both features had similar importance, but the pitch characteristics were more important in the case that the gender of the speakers was different. The least important feature was the energy contour. We have also shown that dual transplantations had complementary scores as expected.

In the second part, we have developed new methods that can be used for voice conversion system design (Chapter 5). Two new vocal tract conversion methods based on subband processing were developed: the DWT based system (Section 5.1) and the selective preemphasis-based system (Section 5.2). We have shown that these methods possess comparable performance at lower prediction orders. This is an advantage for voice conversion applications at high sampling rates. The subband-based framework provides flexibility in voice conversion algorithms because of the following reasons:

- Different frequency bands can be analyzed and modified using different amounts of spectral resolution. Even different methods can be employed at different subbands.
- It is possible to obtain the mapping between the source and the target acoustical spaces.
- The computational load is reduced.

We have also developed a new speaker-specific intonational model in Section 5.4. A segmental pitch contour model for pitch contour transformation. We have shown that this model yields satisfactory results both in pitch contour modeling and transformation. The average similarity to target speaker is increased by employing the segmental model as shown in Section 6.2.2.

In the last part, we have used subjective and objective testing for the evaluation of the new methods (Chapter 6). We have designed a voice conversion database in Turkish. In a subjective test, we have shown that the voice conversion output by the DWT based system was preferred over the output of the full-band based system by 92.1 percent. In this test, we have not employed preemphasis in the full-band based spectral estimation. We have used the subjective testing procedure described in Section 6.2.2 for the evaluation of the new methods proposed and the STASC system. Three vocal tract conversion methods and two pitch transformation strategies were evaluated. The vocal tract conversion methods include the full-band based system in STASC (Section 3.4.2), the DWT based system (Section 5.1), and the selective preemphasis based system (Section 5.2). The mean/variance model (Section 3.6.1) and the segmental pitch contour model (Section 5.4) were used for pitch transformation. We have shown that it is possible to obtain satisfactory performance at lower prediction orders using the DWT based system and the selective preemphasis based system as compared to the full-band based system. The segmental model improved the similarity to target speaker when employed with vocal tract conversion.

In Section 6.3.1, we have shown that the selective preemphasis based spectral estimation performs better than the LP analysis at a lower prediction order. We have compared the performance of the new voice conversion methods using objective tests in Section 6.3.2. The full-band based system performed better than the subband based systems, but the results were closer when pitch transformation is employed.

We are planning to conduct research on the integration of voice conversion and TTS. Several experiments with multilingual synthesizers have shown that the voice conversion methods described in this study can be used for personification of TTS systems. The extensive phonetic and prosodic information in the TTS databases can be used for developing new voice conversion methods suitable for TTS applications.

We have observed that better alignment improves the quality of the voice conversion output considerably. So, we have started collecting a multi-speaker database for accurate phonetic alignment. This method will also enable the use of contextual information in voice conversion. We are expecting to have further improvements by employing contextual information as the short-term characteristics of the speech signals are strongly correlated with the context. We will also be able to use contextual information in the segmental pitch model. In this manner, the pitch segments can be matched and modified applying contextual constraints.

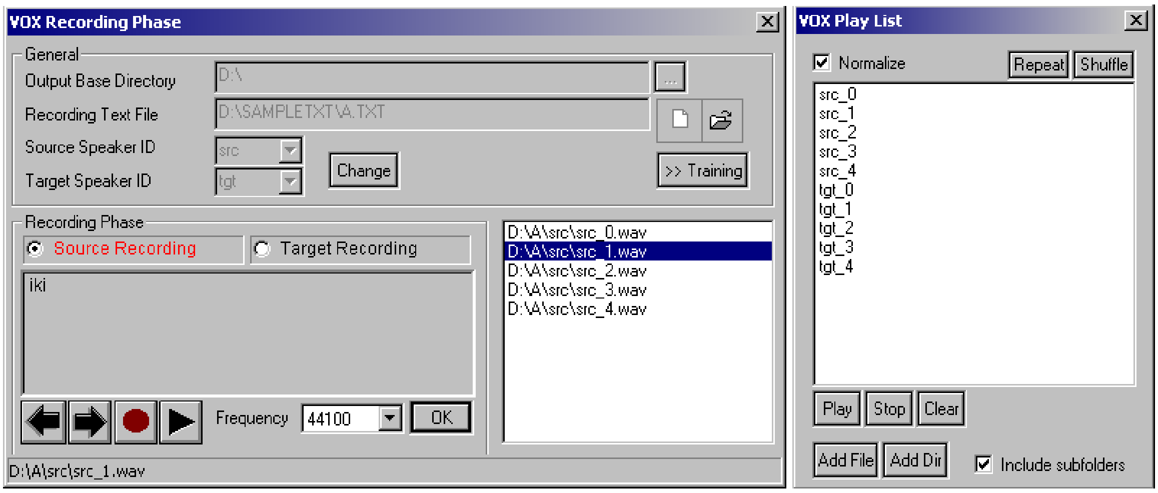
Although the test results obtained during this study provide extensive information on the performance of the new methods proposed, it would be helpful to perform extended subjective tests using more speaker pairs in voice conversion and employing more subjects.

APPENDIX A: VOX – A VOICE CONVERSION SOFTWARE

VOX is a voice conversion software for Windows 98/ME/2000/XP. It has been developed during this study for incorporating all the necessary tools and components of a voice conversion system in a single interface. It is possible to perform fast, efficient, and convincing voice conversion using VOX. It includes the following components:

- Waveform Playing/Recording/Editing Interfaces
- Training and Transformation Interfaces
- Acoustic Feature Transplantation Interface
- Subband Processing Toolbox
- Speech and Audio Processing Tools (PSOLA Interface, Pitch Contour Calculation, Spectral Analysis Tools, Enhancement Interface, Filter Design Toolbox, Equalizer Interface, Reverbarator Interface, Manual Labeling Tools)
- Objective and Subjective Testing Interfaces

Several snapshots from VOX are shown in Figures A.1-A.9. The software was developed in MS Visual C++ 6.0 using the MFC architecture.



VOX Recording Phase

General

Output Base Directory: D:\

Recording Text File: D:\SAMPLETXT\A.TXT

Source Speaker ID: src

Target Speaker ID: tgt

Recording Phase

● Source Recording ○ Target Recording

iki

Frequency: 44100

D:\A\src\src_0.wav
D:\A\src\src_1.wav
D:\A\src\src_2.wav
D:\A\src\src_3.wav
D:\A\src\src_4.wav

D:\A\src\src_1.wav

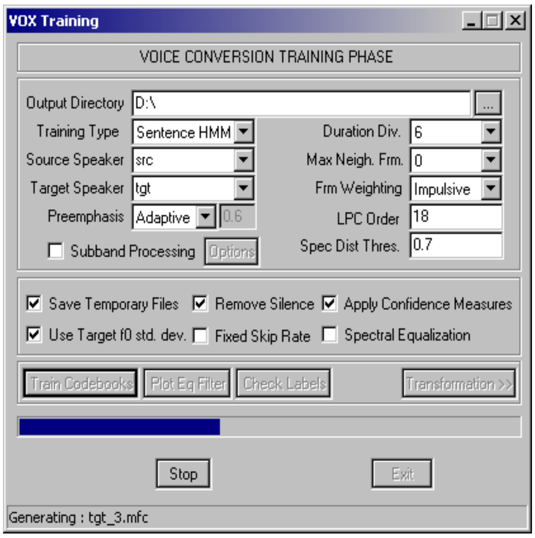
VOX Play List

☑ Normalize

src_0
src_1
src_2
src_3
src_4
tgt_0
tgt_1
tgt_2
tgt_3
tgt_4

☑ Include subfolders

Figure A.1. Audio recording (left), and playing (right) interfaces



VOICE CONVERSION TRAINING PHASE

Output Directory D:\

Training Type Sentence HMM

Source Speaker src

Target Speaker tgt

Preemphasis Adaptive 0.6

□ Subband Processing Options

Duration Div. 6

Max Neigh. Frm. 0

Frm Weighting Impulsive

LPC Order 18

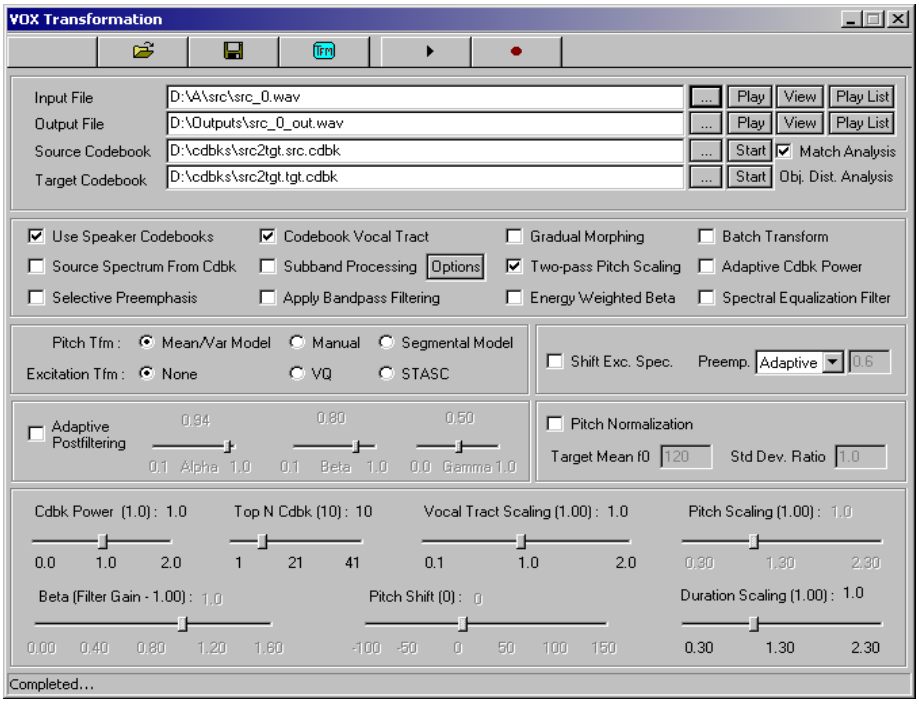
Spec Dist Thres. 0.7

☑ Save Temporary Files ☑ Remove Silence ☑ Apply Confidence Measures

☑ Use Target f0 std. dev. □ Fixed Skip Rate □ Spectral Equalization

Generating : tgt_3.mfc

Figure A.2. Training interface



Input File D:\A\src\src_0.wav

Output File D:\Outputs\src_0_out.wav

Source Codebook D:\cdbks\src2tgt.src.cdbk ☑ Match Analysis

Target Codebook D:\cdbks\src2tgt.tgt.cdbk Obj. Dist. Analysis

☑ Use Speaker Codebooks ☑ Codebook Vocal Tract □ Gradual Morphing □ Batch Transform

□ Source Spectrum From Cdbk □ Subband Processing Options ☑ Two-pass Pitch Scaling □ Adaptive Cdbk Power

□ Selective Preemphasis □ Apply Bandpass Filtering □ Energy Weighted Beta □ Spectral Equalization Filter

Pitch Tfm : ● Mean/Var Model ○ Manual ○ Segmental Model

Excitation Tfm : ● None ○ VQ ○ STASC

□ Shift Exc. Spec. Preemp. Adaptive 0.6

□ Adaptive Postfiltering 0.94 0.1 Alpha 1.0 0.80 0.1 Beta 1.0 0.50 0.0 Gamma 1.0

□ Pitch Normalization

Target Mean f0 120 Std Dev. Ratio 1.0

Cdbk Power (1.0) : 1.0 0.0 1.0 2.0

Top N Cdbk (10) : 10 1 21 41

Vocal Tract Scaling (1.00) : 1.0 0.1 1.0 2.0

Pitch Scaling (1.00) : 1.0 0.30 1.30 2.30

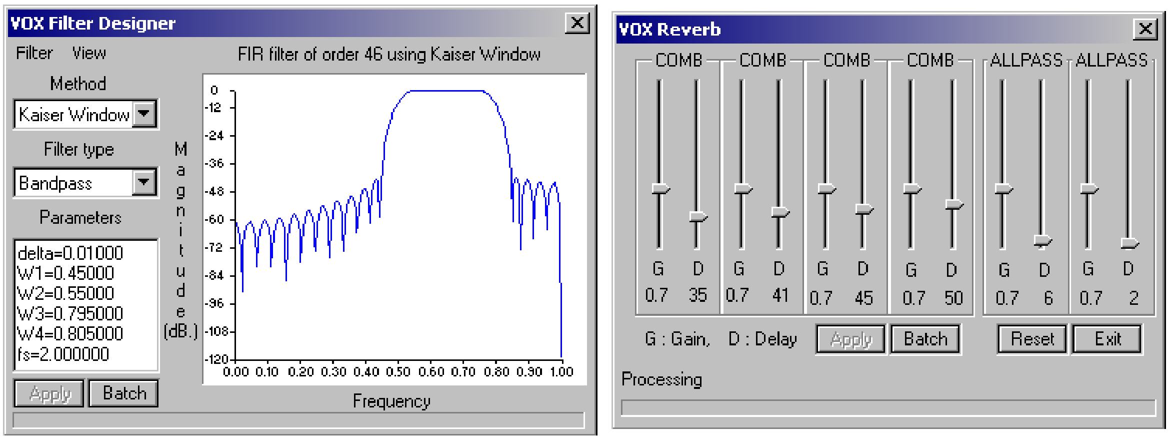
Beta (Filter Gain - 1.00) : 1.0 0.00 0.40 0.80 1.20 1.60

Pitch Shift (0) : 0 -100 -50 0 50 100 150

Duration Scaling (1.00) : 1.0 0.30 1.30 2.30

Completed...

Figure A.3. Transformation interface



VOX Filter Designer

FIR filter of order 46 using Kaiser Window

Method: Kaiser Window

Filter type: Bandpass

Parameters:

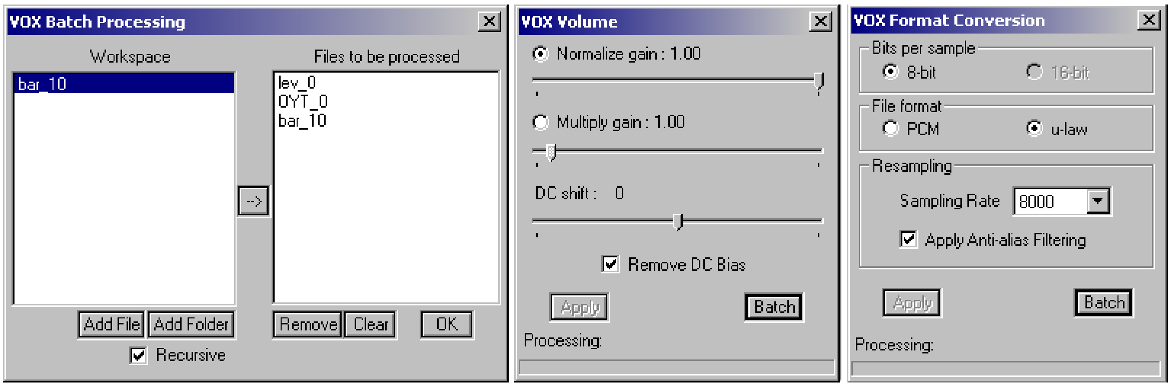
delta=0.01000
W1=0.45000
W2=0.55000
W3=0.795000
W4=0.805000
fs=2.000000

VOX Reverb

COMB		COMB		COMB		COMB		ALLPASS		ALLPASS	
G	D	G	D	G	D	G	D	G	D	G	D
0.7	35	0.7	41	0.7	45	0.7	50	0.7	6	0.7	2

G : Gain, D : Delay

Figure A.4. Filter designer (left) and reverb (right) interfaces



VOX Batch Processing

Workspace: bar_10

Files to be processed:
lev_0
OYT_0
bar_10

☑ Recursive

VOX Volume

● Normalize gain : 1.00

○ Multiply gain : 1.00

DC shift : 0

☑ Remove DC Bias

VOX Format Conversion

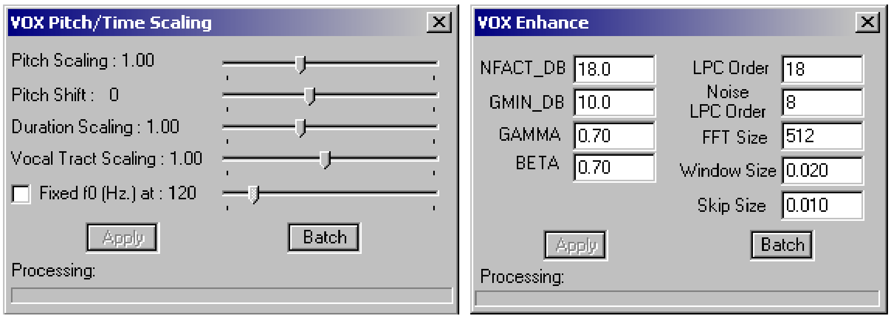
Bits per sample: ● 8-bit ○ 16-bit

File format: ○ PCM ● u-law

Resampling: Sampling Rate 8000

☑ Apply Anti-alias Filtering

Figure A.5. Batch processing (left), volume (middle), and format conversion (right) interfaces



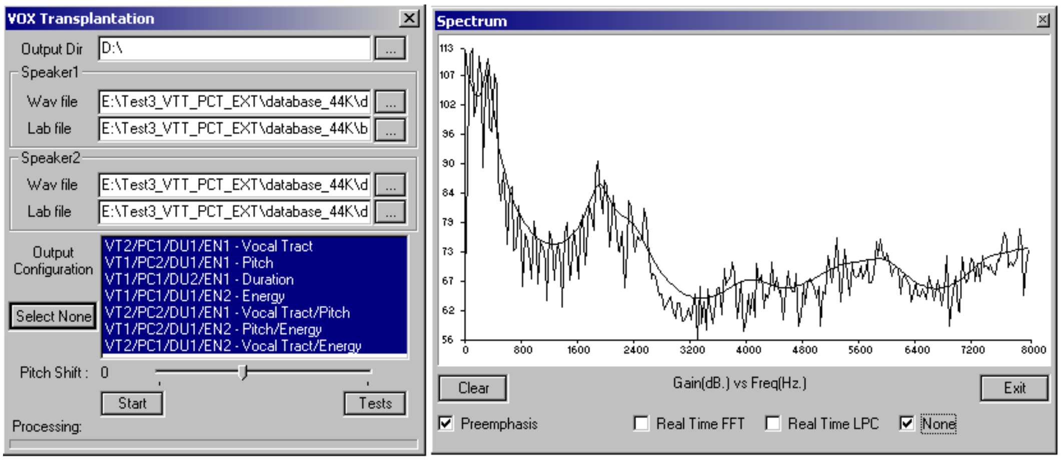
VOX Pitch/Time Scaling

Pitch Scaling : 1.00
Pitch Shift : 0
Duration Scaling : 1.00
Vocal Tract Scaling : 1.00
□ Fixed f0 (Hz.) at : 120

VOX Enhance

NFACT_DB	18.0	LPC Order	18
GMIN_DB	10.0	Noise LPC Order	8
GAMMA	0.70	FFT Size	512
BETA	0.70	Window Size	0.020
		Skip Size	0.010

Figure A.6. PSOLA (left), and enhancement (right) interfaces



Output Dir D:\

Wav file E:\Test3_VTT_PCT_EXT\database_44K\d
Lab file E:\Test3_VTT_PCT_EXT\database_44K\b

Wav file E:\Test3_VTT_PCT_EXT\database_44K\d
Lab file E:\Test3_VTT_PCT_EXT\database_44K\d

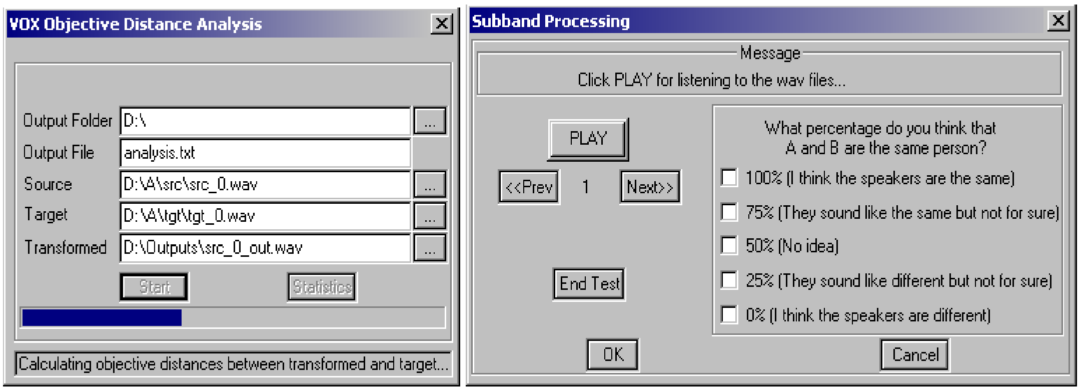
Output Configuration
VT2/PC1/DU1/EN1 - Vocal Tract
VT1/PC2/DU1/EN1 - Pitch
VT1/PC1/DU2/EN1 - Duration
VT1/PC1/DU1/EN2 - Energy
VT2/PC2/DU1/EN1 - Vocal Tract/Pitch
VT1/PC2/DU1/EN2 - Pitch/Energy
VT2/PC1/DU1/EN2 - Vocal Tract/Energy

Pitch Shift : 0

Processing:

☑ Preemphasis ☐ Real Time FFT ☐ Real Time LPC ☑ None

Figure A.7. Acoustic feature transplantation (left) and spectral analysis interfaces



Output Folder D:\
Output File analysis.txt
Source D:\A\src\src_0.wav
Target D:\A\tgt\tgt_0.wav
Transformed D:\Outputs\src_0_out.wav

Calculating objective distances between transformed and target...

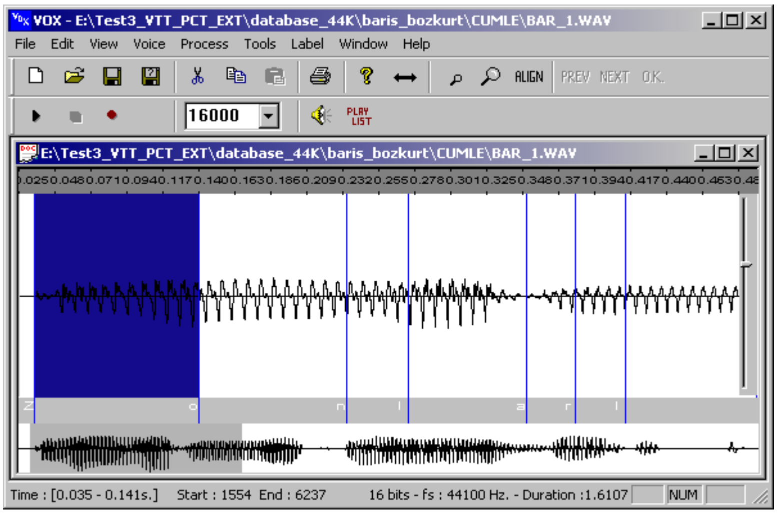
Click PLAY for listening to the wav files...

1

What percentage do you think that
A and B are the same person?

☐ 100% (I think the speakers are the same)
☐ 75% (They sound like the same but not for sure)
☐ 50% (No idea)
☐ 25% (They sound like different but not for sure)
☐ 0% (I think the speakers are different)

Figure A.8. Objective (left) and subjective (right) testing interfaces



VOX - E:\Test3_VTT_PCT_EXT\database_44K\baris_bozkurt\CUMLE\BAR_1.WAV

E:\Test3_VTT_PCT_EXT\database_44K\baris_bozkurt\CUMLE\BAR_1.WAV

Time : [0.035 - 0.141s.] Start : 1554 End : 6237 16 bits - fs : 44100 Hz. - Duration :1.6107

Figure A.9. Waveform editing interface

REFERENCES

Abe, M., S. Nakamura, K. Shikano and H. Kuwabara, 1988, “Voice Conversion Through Vector Quantization”, *Proceedings of the IEEE ICASSP 1988*, pp. 565-568.

Arslan, L. M. and D. Talkin, 1997, “Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum”, *Proceedings of the EUROSPEECH 1997*, Rhodes, Greece, Vol. 3, pp. 1347-1350.

Arslan, L.M., 1999, “Speaker Transformation Algorithm Using Segmental Codebooks”, *Speech Communication* 28 (1999), pp. 211-226.

Baudoin, G. and Y. Stylianou, 1996 “On the transformation of the speech spectrum for voice conversion”, *Proceedings of the ICSLP 1996*, Philedelphia, USA, pp. 1405 – 1408, 1996.

Burrus, C. S., R. A. Gopinath and H. Guo, 1998, *Introduction to Wavelets and Wavelet Transforms*, Prentice-Hall Inc., New Jersey.

Childers, D.G., 1995, “Glottal Source Modeling for Voice Conversion”, *Speech Communication* 16 (2) (1995), pp. 127-138.

Chappel, D.T. and J.H.L. Hansen, 1998, "Speaker-Specific Pitch Contour Modeling and Modification", *Proceedings of the IEEE ICASSP*, Seattle, Washington, May 1998, Vol. II, pp. 885-888.

Drioli, C., 1999, “Radial Basis Function Networks For Conversion of Sound Spectra”, *Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99)*, NTNU, Trondheim, December 9-11, 1999.

Fant, G., J. Liljencrants and Q. Lin, 1985, “A Four-Parameter Model of the Glottal Flow”, *Speech Transmission Laboratory Quarterly Progress and Status Reports*, No. 4, Royal Institute of Technology, Stockholm, Sweden, 1-13, 1985.

Flanagan, J. L. and R. M. Golden, 1966, “Phase Vocoder”, *Bell Systems Technical Journal,* Vol. 45, pp. 1493-1509, 1966.

Fujisaki, H. and H. Kawai, 1982, “Modeling the Dynamic Characteristics of Voice Fundamental Frequency with Applications to Analysis and Synthesis of Intonation”, *Working Group on Intonation, 13th International Congress of Linguists*, Tokyo.

Gold, B. and L. R. Rabiner, 1969, “Parallel Processing Techniques for Estimating Pitch periods of Speech in the Time Domain”, *Journal of the Acoustical Society of America*, Vol. 46, No. 2, Pt. 2, pp. 442-448, August 1969.

Griffin, D.W., 1987, *Multi-Band Excitation Vocoder*, Ph.D. Dissertation, Massachusetts Institute Of Technology.

Gutierrez-Arriola, J. M., Y. S. Hsiao, J. M. Montero, J.M. Pardo and D. G. Childers, 1998, “Voice Conversion Based On Parameter Transformation”, *Proceedings of the ICSLP 1998*, Vol. 3, pp. 987-990, 30 Nov-4 Dec. 1998, Sydney, Australia.

Hardwick, J. C. and J. S. Lim, 1988, “A 4.8 Kbps Multi-Band Excitation Speech Coder”, *Proceedings of the IEEE ICASSP 1988*, pp. 374-377.

Hashimoto, M. and N. Higuchi, 1996, “Training Data Selection for Voice Conversion Using Speaker Selection and Vector Field Smoothing”, *Proceedings of the ICSLP 1996*, pp. 1397-1400.

ISO/IEC, 1993, “Information technology - Coding of moving pictures and associated audio for digital storage media at up to 1,5 Mbits/s - Part3: audio”, *British standard*, October 1993, Implementation of ISO/IEC 11172-3:1993, BSI, London, First edition 1993-08-01.

Itakura, F., 1975, "Minimum Prediction Residual Principle Applied to Speech Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-23, No. 1, pp. 67-72, February, 1975.

Kain, A. and M. Macon, 1998a, "Personalizing A Speech Synthesizer by Voice Adaptation", *Proceedings of the Third ESCA/COCOSDA International Speech Synthesis Workshop*, November 1998, pp. 225-230.

Kain, A. and M. Macon, 1998b, "Spectral Voice Conversion for Text-To-Speech Synthesis", *Proceedings of the IEEE ICASSP 1998*, May 1998, Vol. 1, pp. 285-288.

Kain, A. and M. Macon, 1998c, "Text-To-Speech Voice Adaptation from Sparse Training Data", *Proceedings of the ICSLP 1998*, November 1998, vol. 7, pp. 2847-50.

Kain, A. and Y. Stylianou, 2000, "Stochastic Modeling Of Spectral Adjustment for High Quality Pitch Modification", *Proceedings of the IEEE ICASSP 2000*, Vol. 2, pp. 949-952, June 2000, Istanbul, Turkey.

Kain, A. and M. W. Macon, 2001, "Design and Evaluation of a Voice Conversion Algorithm Based On Spectral Envelope Mapping And Residual Prediction", *Proceedings of the IEEE ICASSP 2001*, May 2001, Salt Lake City, Utah, USA.

Kain, A. B., 2001, *High Resolution Voice Transformation*, Ph.D. Dissertation, OGI School of Science and Engineering at Oregon Health and Science University.

Kenny, T.W., 2001, *ME 117/220: Introduction to Sensors Lecture Notes*, http://design.stanford.edu/Courses/me220/lectures/lect01/auditory.html

Kim, E. K., S. Lee and Y. H. Oh, 1997, “Hidden Markov Model Based Voice Conversion Using Dynamic Characteristics of Speaker”, *Proceedings of the EUROSPEECH 1997*, Rhodes, Greece.

Laroche, J. and M. Dolson, 1999, "New Phase Vocoder Technique for Pitch-Shifting, Harmonizing and Other Exotic Effects", *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, New Paltz, NY., 1999.

McAulay, R. J. and T. F. Quatieri, 1995, “Sinusoidal Coding” in Kleijn and Paliwal (eds.), *Speech Coding And Synthesis*, pp. 121-173, Elsevier Science B.V., Netherlands.

Moulines, E. and F. Charpentier, 1990, “Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis Using Diphones”, *Speech Communication* 9 (1990) pp. 453-467.

Moulines, E. and W. Verhelst, 1995, “Time-Domain and Frequency-Domain Techniques for Prosodic Modification of Speech” in Kleijn and Paliwal (eds.), *Speech Coding And Synthesis*, pp. 519-555, Elsevier Science B.V., Netherlands.

Narusawa, S., N. Minematsu, K. Hirose and H. Fujisaki, 2002, “Automatic Extraction of Model Parameters From Fundamental Frequency Contours of English Utterances”, *Proceedings. of the ICSLP 2002*, Vol. 3, pp. 1725-1728, September 2002, Denver, Colorado, USA.

Nishiguchi, M., J. Matsumoto, R. Wakatsuki and S. Ono, 1993, “Vector Quantized MBE With Simplified V/UV Division At 3.0 Kbps”, *Proceedings of the ICASSP 1993*, pp. II- 151- 154, Apr. 1993.

Ormancı, E., U. H. Nikbay, O. Turk and L. M. Arslan, 2002, “Subjective Assessment of Frequency Bands for Perception of Speaker Identity”, *Proceedings of the ICSLP 2002*, Vol. 4, pp.2581-2584, September 2002, Denver, Colorado, USA.

Pellom, B. L. and J. H. L. Hansen, 1997, “Spectral Normalization Employing Hidden Markov Modeling Techniques of Line Spectrum Pair Frequencies”, *Proceedings of the ICASSP 1997*, Vol 2, pp. 943-946.

Quatieri, T. F. and R. J. McAulay, 1992, “Shape Invariant Time-Scale and Pitch Modification of Speech”, *IEEE Transactions On Signal Processing*, Vol. 40, No. 3, March 1992, pp. 497-510.

Rabiner, L. R. and R. W. Schafer, 1978, *Digital Processing of Speech Signals*, Prentice-Hall Inc., Englewood Cliffs. New Jersey.

Ribeiro, C. M. and I. M. Trancoso, 1997, “Phonetic Vocoding with Speaker Adaptation”, *Proceedings of the Eurospeech 1997*, Rhodes, Greece.

Rinscheid, A., 1996, “Voice Conversion Based on Topological Feature Maps and Time-Variant Filtering”, *Proceedings of the ICSLP 1996*, October 3-6, 1996, Philedelphia, USA, Vol 3, pp. 1445-1448.

Rothweiler, J., 1999, "A Root-finding Algorithm for Line Spectral Frequencies", *Proceedings of the IEEE ICASSP 1999*, March 15-19, 1999, Phoenix, AZ, USA, pp. II-661 - II-664.

Stylianou, Y., O. Cappe and E. Moulines, 1998, “Continuous Probabilistic Transform for Voice Conversion”, *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 2, March 1998, pp. 131-142.

Stylianou, Y. and O. Cappe, 1998, “A System for Voice Conversion Based On Probabilistic Classification and a Harmonic Plus Noise Model”, *Proceedings of the IEEE ICASSP 1998*, Seattle, Washington, USA, pp. 281-284.

Stylianou, Y., O. Cappe and E. Moulines, 1995, “Statistical Methods for Voice Quality Transformation”, *Proceedings of the Eurospeech 1995*, Madrid, Spain.

Strik, H., 1998, “Automatic Parametrization of Differentiated Glottal Flow: Comparing Methods by means of Synthetic Flow Pulses”, *Journal of the Acoustical Society of America*, Vol. 103 (5), pt. 1, May 1998, pp. 2659-2669.

Syrdal, A. K., A. Conkie and Y. Stylianou, 1998, “Exploration of Acoustic Correlates in Speaker Selection for Concatenative Synthesis”, *Proceedings of the ICSLP 1998*, November 1998, Syndey, Australia.

Talkin, D., 1995, “A Robust Algorithm for Pitch Tracking (RAPT)”, in Kleijn and Paliwal (eds.), *Speech Coding And Synthesis*, pp. 121-173, Elsevier Science B.V., Netherlands.

Tang, M., C. Wang and S. Seneff, 2001, “Voice Transformations: From Speech Synthesis to Mammalian Vocalizations”, *Proceedings of the Eurospeech 2001*, Aalborg, Denmark, September 2001.

Tanaka, K. and M. Abe, 1997, “A New Fundamental Frequency Modification Algorithm with Transformation of Spectrum Envelope According to f0”, ICASSP 1997, Vol. 2, pp. 951-954.

Torresani, B., 1999, “An Overview of Wavelet Analysis and Time-Frequency Analysis (A Minicourse)”, in V.B. Priezzhev and V.P. Spiridonov (eds.), *Self-Similar Systems*, *Proceedings of the International Workshop 1998*, JINR, E5-99-38, Dubna, 1999, 9-34.

Turk, O. and L. M. Arslan, 2002, “Subband Based Voice Conversion”, *Proceedings of the ICSLP 2002*, Vol. 1, pp.289-292, September 2002, Denver, Colorado, USA.

Verhelst, W., T. Ceyssens and P. Wambacq, 2002, “On Inter-Signal Transplantation Of Voice Characteristics”, *Proceedings of the 3rd IEEE Benelux Signal Processing Symposium (SPS 2002)*, Leuven, Belgium, pp. 137-140.

Violaro, F. and O. Böeffard, 1998, “ A Hybrid Model for Text-to-Speech Synthesis”, *IEEE Transactions On Speech and Audio Processing*, Vol.6, No. 5, September 1998.

Watanabe, T., T. Murakami, M. Namba, T. Hoya and Y. Ishida, 2002, “Transformation of Spectral Envelope For Voice Conversion Based On Radial Basis Function Networks”, *Proceedings of the ICSLP 2002*, Vol. 1, pp. 285-288.

Wang, T., K. Tang and C. Feng, 1996, “A High Quality MBE-LPC-FE Speech Coder At 2.4 Kbps and 1.2 Kbps”, *Proceedings of the IEEE ICASSP 1996*, pp. 208-211.

Wu, M., 2003, *ENEE 408G Multimedia Signal Processing Lecture Notes*, http://www. enee.umd.edu/class/enee408g/EE408G_S03_schedule.html

Yapanel, Ü., 2000, “Garbage Modeling Techniques For A Turkish Keyword Spotting System”, M.S. Thesis, Boğaziçi University.